Towards a semantic granularity model for domain-specific information retrieval

Towards a Semantic Granularity Model for Domain-specific Information Retrieval

XIN YAN, University of Queensland
RAYMOND Y.K. LAU, City University of Hong Kong
DAWEI SONG, The Robert Gordon University
XUE LI, University of Queensland
JIAN MA, City University of Hong Kong

Both similarity-based and popularity-based document ranking functions have been successfully applied to information retrieval (IR) in general. However, the dimension of semantic granularity also should be considered for effective retrieval. In this paper, we propose a semantic granularity based IR model which takes into account the three dimensions, namely similarity, popularity, and semantic granularity, to improve domain-specific search. In particular, a concept-based computational model is developed to estimate the semantic granularity of documents with reference to a domain ontology. Semantic granularity refers to the levels of semantic detail carried by an information item. The results of our benchmark experiments confirm that the proposed semantic granularity based IR model performs significantly better than the similarity-based baseline in both a bio-medical and an agricultural domain. In addition, a series of user-oriented studies reveal that the proposed document ranking functions resemble the implicit ranking functions exercised by humans. The perceived relevance of the documents delivered by the granularity-based IR system is significantly higher than that produced by a popular search engine for a number of domain-specific search tasks. To the best of our knowledge, this is the first study regarding the application of semantic granularity to enhance domain-specific IR.

Categories and Subject Descriptors:
H.3.1 [**Content Analysis and Indexing**]: Abstracting Methods, Dictionaries, Indexing Methods; **H.3.3** [**Information Search and Retrieval**]: Retrieval Models, Search Process; **H.3.4** [**Systems and Software**]: performance evaluation (efficiency and effectiveness); **H.3.m** [**Miscellaneous**]: Theoretical Study of Information Retrieval
General Terms: Theory, Algorithms, Experimentation
Additional Key Words and Phrases: Document Ranking, Domain-specific Search, Domain Ontology, Information Retrieval, Granular Computing.

This work is supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) Research Grant EP/F035705/1, the City University of Hong Kong's Start-up Grant 7200126, and the City University of Hong Kong's Strategic Research Grant 7002426.
Author's addresses: X. Yan and X. Li, School of Information Technology and Electrical Engineering, University of Queensland; R.Y.K. Lau and J. Ma, Department of Information Systems, City University of Hong Kong; D. Song, School of Computing, The Robert Gordon University.

DOI10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The growing amount and diversity of information archived on electronic networks, such as the Web, is making it increasingly difficult for information seekers to locate relevant information [Lau et al. 2008; Ho and Tang 2001]. Classical similarity-based IR models and the more recent popularity-based IR models have successfully supported IR in general and Web searching in particular. However, similarity and popularity based IR models alone might not be effective enough to support domain-specific IR. The possible weaknesses of these models for domain-specific IR will be discussed in Section 1.1. In this paper, we develop an effective IR model for domain-specific searching by exploring a new dimension of IR called "semantic granularity" to supplement the well-known notions of "similarity" and "popularity". The term "information granularity" is not new to the research community investigating granular computing, which uses levels of "granularity" or "abstraction" to systematically represent, analyze, and solve real-world problems [Bargiela and Pedrycz 2008; Yao 2005]. "Information granulation" refers to the computational processes of generating and presenting levels of abstraction to facilitate problem solving [Yao 2005; Zadeh 1979]. Within the field of granular computing, information granularity usually refers to "structural granularity", which signifies the structural abstraction of information items. A structural abstraction can be based on a complete information item, such as a document, or its constituent parts, such as the sentences. For instance, the structural abstractions of a book can be generated from chapters, sections, pages, paragraphs, and so on. Previous IR research on structural granularity has been conducted under the headings of passage retrieval [Liu and Croft 2002; Wang and Si 2008] and entity-based searches, such as expert search [Bailey et al. 2007]. However, this paper will focus on the relatively new area of "semantic granularity", which has received little attention in IR research to date. Semantic granularity refers to the levels of semantic detail carried by an information item [Fonseca et al. 2002]. In this paper, the term "granularity" signifies semantic rather than structural granularity.

Existing Web-based IR systems such as Google Maps[1] perform information granulation for special kinds of information (i.e., geographical maps). As shown in Figure 1, the slider bar (a form of granularity control) allows information seekers to view geographical locations at different levels of granularity. For example, the information seeker can use the granularity control to obtain general information concerning the location of the ACM headquarters (Figure 1a) or more specific details about the subway stations close to the ACM head office (Figure 1b). Nevertheless, existing information granulation mechanisms do not effectively support Web document searching as accurately estimating the semantic details carried by such documents remains an extremely challenging

[1] http://maps.google.com/

problem.

a. A High Level View of the ACM Headquarters

b. A Finer View of the ACM Headquarters

Figure 1. An Example of Information Granulation

1.1 The Needs for Granularity-Based IR

To verify that "similarity" and "granularity" are two fundamentally different dimensions of IR, a pilot study was conducted on the collection of bio-medical documents contained in the OHSUMED corpus[2] [Hersh et al. 1994]. For every test query to the OHSUMED benchmark collection, two sets of document scores were computed using the similarity and granularity based document ranking functions, respectively. The computational details of the granularity-based document ranking functions are illustrated in Section 3.4. The similarity-based document ranking scores were calculated using Lucene[3], a well-known open source IR system. Our empirical results revealed a very low degree of correlation between these two document-ranking functions. Diagrams plotting the correlations between the two sets of document scores that were generated by the respective document ranking functions are illustrated in Appendix A.

Our results show that there is an obvious need for granularity based IR for specific domains. In addition, we provide the following examples to motivate the development of granularity-based IR for domain-specific searches. For example, while medical professionals tend to search for specific technical articles relating to particular medical topics, the general public may wish to retrieve general information about a disease or a medicine. If an information seeker were to search for "general AIDS information" via a

[2] http://ir.ohsu.edu/ohsumed/

[3] http://lucene.apache.org/java/docs/index.html

medical search tool, such as PubMed[4], thousands of documents regarding different aspects of AIDS, such as treatment, drug therapy, transmission, diagnosis, and history, would be retrieved. This presents a challenge for traditional similarity-based IR systems, as similarity computation based on keywords cannot always distinguish between general and specific documents. For instance, based on a similarity-based document ranking function, the above search may rank a specific research paper, such as "*Multiple Dimensions of HIV Stigma and Psychological Distress Among Asians and Pacific Islanders*", higher than the general AIDS publication, "*HIV/AIDS: A Minority Health Issue*". In this case, however, the lower ranked paper is the most relevant to the information seeker's request for "general AIDS information".

As a further example of the need for granularity-based IR, suppose your horse is sick and you want to find out why. You then send a query about the "causes of African Horse Sickness (AHS)" to the Google search engine, which employs a combined similarity and popularity based document ranking function. As shown in Figure 2, the first document Google returns is the AHS Facebook community page,[5] which describes a Facebook charity network (i.e., a "cause") relating to AHS, rather an explanation of the possible causes of the AHS disease. Although the AHS Facebook community page is very popular (probably read by many Facebook users and generating many in-links), it is not relevant to your specific query. As this example suggests, a document ranking function which only considers similarity and popularity might not provide effective support for domain-specific IR.

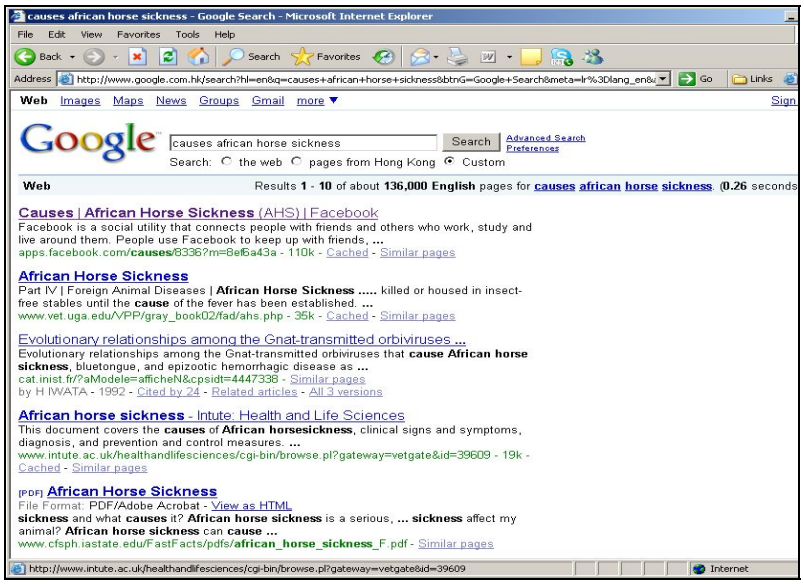

Fig. 2. Top Query Results for the AHS Query

[4] http://www.ncbi.nlm.nih.gov/pubmed/

[5] http://apps.facebook.com/causes/8336?facebook_url=true as accessed on 19 April 2009

Because of the sheer volume of documents archived on the Web and the growing number of domain-specific document repositories, it is extremely difficult, if not impossible, to manually label "general" or "specific" documents. Accordingly, there is a pressing need for an alternative computational model capable of facilitating domain-specific IR. Granularity-based IR (or granular IR) aims to find documents that are not only similar to a query but which also satisfy a specific granularity requirement defined in relation to a wide granularity spectrum of semantically general and semantically specific information [Lau et al. 2009b]. Because generality is the antonym of specificity, we can estimate the granularity (i.e., an attribute or property) of a document in terms of its informational generality (i.e., an attribute value).

1.2 A Granular IR Model for Domain-specific Search

An effective document ranking function is essential for a successful IR system as information seekers rarely review documents beyond the first page of a result set [Granka et al. 2004]. According to the probabilistic ranking principle, an IR system that ranks retrieved documents in an order of decreasing probability of relevance to a query can improve IR performance [Robertson 1997]. In general, document relevance can be estimated using a variety of similarity functions. For example, the similarity between a document and a query can be estimated by measuring the cosine angle between the corresponding vectors in a vector space [Salton et al. 1975]. Alternatively, similarity can be computed in a probabilistic sense by estimating the likelihood of a document generating a particular query [Ponte and Croft 1998].

Following the invention of the PageRank algorithm and its variants [Haveliwala 2003; Page et al. 1998], popular Internet search engines have employed hybrid similarity-based and popularity-based mechanisms to rank Web documents. Popularity-based ranking functions implicitly assume that popularity closely correlates with relevance. Unfortunately, the correlation between popularity and relevance could be weak for newly created Web pages with few inlinks [Mowshowitz and Kawaguchi 2002]. To improve the effectiveness of domain-specific IR, we propose a novel semantic IR model which can take into account three important dimensions of IR, namely "similarity", "popularity", and "granularity".

In practice, there are three possible ways of determining the granularity requirements of an information seeker: (1) manual detection of query granularity – an information seeker explicitly specifies granularity by labeling a query general or specific to indicate whether general or specific documents are required; (2) semi-automatic detection of query granularity – an information seeker uses a set of pre-defined words, such as "review", "introduction", "in-depth", "specialized", etc. to specify their granularity preferences; (3)

automatic detection of query granularity – an IR system automatically estimates the granularity requirements of a query using the same approach for document granularity computation. In this paper, we assume that the information granularity an information seeker requires can be identified through any one of the above three ways. Figure 3 shows a snapshot view of the main interface of our proposed granular IR system. The granularity control bar allows users to manually specify their granularity preferences. Figure 4 shows the result window of the granular IR system after applying granularity-based document ranking to the query "causes African Horse Sickness". The computational details of the granularity-based document ranking functions are illustrated in Section 3.4.

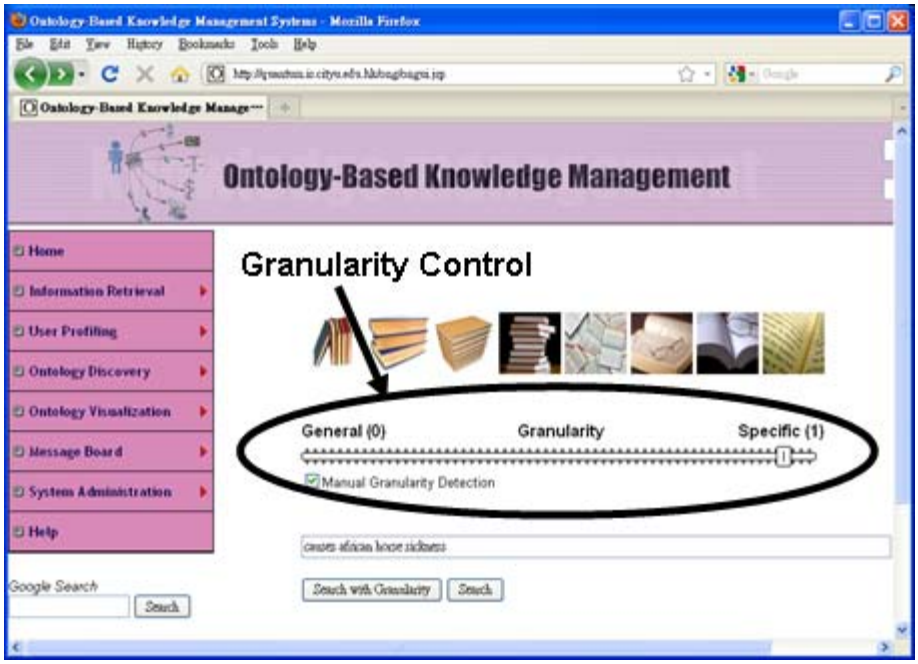

Fig. 3. A Snapshot View of the Granular IR System

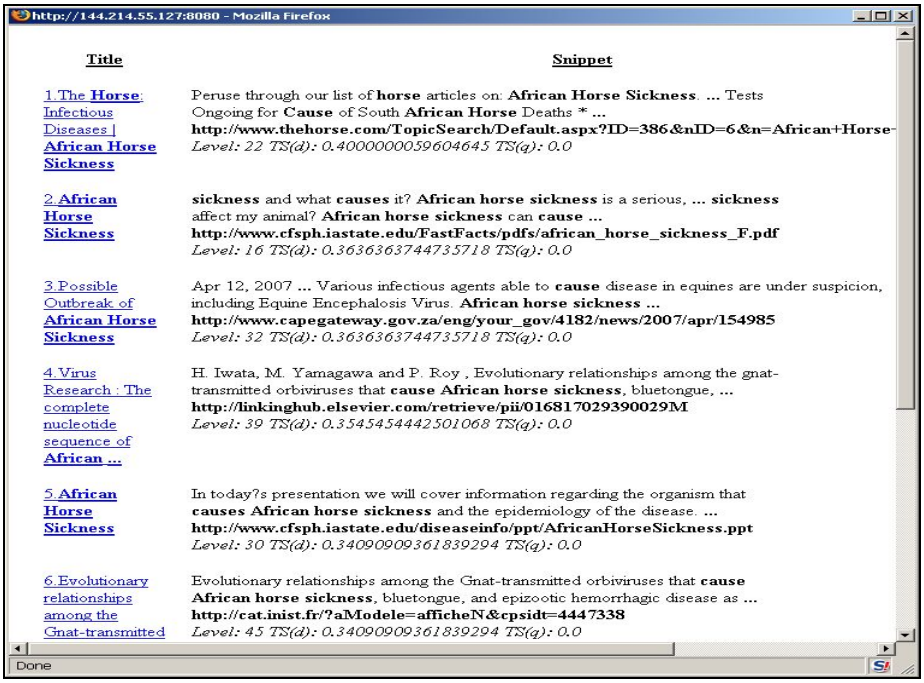

Fig. 4. The Result Window of the Granular IR System

1.3 Contributions of the Paper

The main theoretical contributions of the research presented in this paper are: (1) the design of a novel granular IR model, based on the three dimensions of similarity, popularity, and granularity, to improve the effectiveness of domain-specific IR; (2) the development of an ontology-based computational model that estimates document (query) granularity by analyzing the semantic contents captured in a document (query); (3) the development of a novel concept marking method to efficiently identify domain-specific concepts presented in a document; (4) the empirical validation of the proposed granular IR model based on a series of benchmark tests and usability studies. Our research has practical implications as the development of an effective granular IR model will facilitate domain-specific searching and, thereby, help alleviate the broader problems associated with information overload.

1.4 Structure of the Paper

The rest of this paper is organized as follows. Section 2 reviews the literature relating to granular computing and information retrieval, and positions our work among the existing research. Section 3 illustrates our proposed computational models for estimating document granularity and re-ranking documents based on semantic granularity. Our system-oriented and user-oriented experiments and the corresponding results are discussed in Section 4. The final section offers concluding remarks and outlines the future directions of our research.

2. RELATED RESEARCH

As granular computing is a relatively new area of research, few studies have explored granular IR. The research closest to our work includes studies on "granular IR support systems", "semantic information relatedness", "aspect retrieval", "query generality", and "subtopic retrieval". There are some previous studies examining text familiarity and text readability. However, this paper does not explore these ideas and the work related to text familiarity and text readability will not be discussed here.

2.1 Granular IR Support Systems

As granular computing is an emerging field of research [Bargiela and Pedrycz 2008; Yao 2005], few studies have examined the design and development of granular IR systems. Yao [Yao 2002] was probably the first to explore the idea of granular computing in the context of IR. He proposed that, to develop effective IR systems for individuals or groups of information seekers, IR support systems would need to exploit document space granulations (such as document clustering), user space granulations (such as grouping similar queries into a user profile), term space granulations (such as grouping terms by specificity or generality), and retrieval result granulations (such as clustering result sets) [Yao 2002]. However, the application of granular computing methodologies to IR has remained at a conceptual level and has yet to be applied to concrete system design and implementation. Our research extends the idea of the granular IR support system to the design, implementation, and evaluation of a prototype granular IR system. In particular, we exploit term space granulation to construct a computational model to estimate the granularity of documents and queries.

Numerous researchers in the field of IR have examined document and result sets clustering (i.e., retrieval result granulation) [Roussinov and Chen 2001; Buyukkokten et al. 2002]. In one study, for instance, different textual units of Web documents were identified, grouped, and summarized to produce satisfactory displays on small handheld devices [Buyukkokten et al. 2002]. As a result, users would be able to quickly access the most important information on a Web page, even though the physical display size of handheld devices is quite limited. In addition to dividing documents into clusters (e.g., specific vs. general), our granular IR system also can assess the semantic granularity of each individual document or query.

2.2 Semantic Information Relatedness

Resnik [Resnik 1995] proposed an information theoretic method of measuring the semantic similarity between pairs of concepts. This hybrid approach combined corpus-based statistical methods with knowledge-based ontological structures. In particular, the semantic similarity of a pair of concepts was derived from the information

content of the least common subsumer (lcs) that subsumed both of these concepts. The information content of the lcs was estimated by comparing the occurrence frequency of all the terms subsumed by the lcs to the occurrence frequency of all the terms encoded in an ontology (i.e., a taxonomic structure). The occurrence frequency of a term was obtained from the actual term occurrence statistics of a reference corpus. An empirical study revealed that when compared to the edge counting method, the information theoretic approach was closer to human judgment for a set of term pairs [Resnik 1995]. Concept similarity may be able to estimate the semantic granularity of a document. For instance, if all the terms of a document are semantically related, the document is considered to be specific to a particular topic.

Allen and Wu measured document generality based on the mean generality of domain concepts contained in a document [Allen and Wu 2002]. In particular, 64 seed terms were pre-defined to form a reference set. Half of these seed terms were considered to be general and the other half specific. It was assumed that the general terms would be more related to each other than the specific terms. The generality of the document could be derived by computing the relatedness between the terms extracted from a document and the pre-defined reference terms. However, constructing a pre-defined set of reference terms for each information domain is very labor-intensive. Moreover, the assumption that general terms are more related to each other than specific terms may not hold, especially for domain-specific IR. For example, in the medical domain, the name of a specific medicine or virus is often closely related to the name of a specific disease. To address these problems, we propose the application of the domain-specific knowledge encoded in a domain ontology to estimate semantic granularity of documents. Semantic granularity can be computed from the terms extracted from a document and their level of generalization with respect to the domain ontology. In particular, conceptual generalization can be estimated according to certain semantic relations, such as the "hypernym" relation encoded in the concept hierarchies. For example, as "illness" is a hypernym of "flu", "illness" is considered to be a more general concept.

2.3 Aspect Retrieval

"Aspect Retrieval" was explored in the interactive track of TREC-6 [Lagergren and Over 1998; Over 1997; Swan and Allan 1998], TREC-7 [Belkin et al. 1998; Bodner and Chignell 1998; Fuller et al. 1998; Gey et al. 1998; Hersh et al. 1998; Ogden et al. 1998; Over 1998; Robertson et al. 1998; Yang et al. 1998], and TREC-8 [Beaulieu et al. 1999; Belkin et al. 1999; Fuller et al. 1999; Hersh 1999; Hersh et al. 1999; Larson 1999; Yang et al. 1999]. Aspect retrieval helps information seekers retrieve documents covering as many different aspects of an information topic as possible given a limited time span. An *aspect* is defined as one of the many possible answers to an information topic [Over 1998; Swan and Allan 1998]. For every document in a collection, the corresponding aspects of

topics are judged by human assessors to create the correct answers and to evaluate the performance of various aspect retrieval methods. Existing research work in the sub-field of aspect retrieval mainly focuses on studying the search behavior of information seekers and user interfaces. To a certain degree, aspect is related to granularity because a document covering many aspects tends to be more general than another document carrying specific aspects of information.

The recent studies of semantic components and their applications to IR [Price et al. 2007] can be regarded as a kind of "Aspect Retrieval". A semantic component captures some metadata about the aspects of a document segment. It is considered to be a complementary mechanism combining full text and keyword indexing [Price et al. 2007]. The instances of a class defined by the semantic component are the text segments semantically related to the "aspect" of that class. In addition, the concepts (classes) defined by the semantic component can be used to construct a semantically rich query to improve retrieval effectiveness. The original motivation of "Aspect Retrieval" was to retrieve documents covering a variety of aspects of a topic. In contrast, the semantic component approach aims to locate specific information items relating to a particular aspect. Even so, labeling (indexing) the text segments with respect to the semantic component required intensive manual effort. Instead of attempting to construct a semantic component, the computational method illustrated in this paper estimates the semantic granularity of documents with reference to an existing domain ontology (i.e., a semantic component). According to the granularity requirement implicitly attached to a query and the estimated semantic granularity of documents, our granularity-based document ranking function can improve the ranking of documents with respect to the particular IR scenario. No human effort is needed to index documents with reference to their granularity.

2.4 Subtopic Retrieval

The development of an automatic granularity-based document ranking function has been implicitly examined in relation to subtopic retrieval [Zhai et al. 2003]. It was argued that, in cases such as a literature survey, documents need to be found that cover as many different subtopics of a general topic as possible [Zhai et al. 2003]. Given a set of documents, a subtopic retrieval method re-ranks the documents according to their generality and relevance with respect to the given query. Statistical language models and maximal marginal relevance were examined in relation to subtopic retrieval [Carbonell and Goldstain 1998].

Another area of research termed "affinity rank" tackled the ranking problem in much the same manner as subtopic retrieval [Liu et al. 2004]. Affinity rank computation is underpinned by two intuitions: (1) the more neighbors a document has, the more informative the document will be; (2) a document is considered informative if its

neighbors also are informative. Information richness was estimated by computing the principal eigenvector of a matrix where each entry represented the similarity value between a pair of documents in a document vector space [Liu et al. 2004]. Affinity ranking and subtopic retrieval are both based on statistical approaches. However, we propose that the granularity of a document is evaluated according to its semantic contents, such as the domain concepts contained in the document. For example, if a document contains general terms with reference to a domain ontology, it is considered a general document.

2.5 Query Generality

There are a number of proposals of how to define query generality in the literature [He and Ounis 2004; Plachouras et al. 2003; Van Rijsbergen 1979]. Van Rijsbergen treated query generality as a measure of the density of relevant documents in a collection [Plachouras et al. 2003; Van Rijsbergen 1979]. Based on van Rijsbergen's proposal, He and Ounis [He and Ounis 2004] defined query specificity ω (an antonym of generality) by: $\omega = -log(\frac{N_Q}{N})$, where N_Q is the total number of documents containing at least one query term and N is the total number of documents in the collection. Based on the above definition, the more documents a query retrieves, the more general (or less specific) the query will be. However, estimating query generality solely based on query term popularity may not be sufficiently accurate. Consider two queries, Q_1 ("AIDS review") and Q_2 ("SARS review"), in the PubMed collection: Q_1 results in 19,311 documents but Q_2 only returns 396 documents. We assume that the size N of the PubMed collection is 11,000,000. According to the aforementioned query specificity measure, the specificity of Q_1 and Q_2 are 6.3450 and 10.2320 respectively. However, it is most likely incorrect to say that Q_1 is more general than Q_2 because both of them are general queries. As "SARS" is a recently discovered disease, we can expect that there will be fewer documents relating to "SARS" in the PubMed collection than there are on "AIDS". Above all, the query specificity measure defined above cannot be computed prior to the actual execution of the query and it does not help to predict query granularity a priori.

3. A COMPUTATIONAL MODEL FOR MEASURING SEMANTIC GRANULARITY

In this section, we first explain the intuition behind our granular IR model with reference to a bio-medical domain. Then, we illustrate the computational details of the granular IR model.

3.1 The Notion of Document Granularity

The notion of document granularity refers to the levels of semantic generality or specificity conveyed by documents. We believe that document granularity can be computed with reference to a domain ontology such as MeSH[6]. Given the fact that

[6]http://www.nlm.nih.gov/mesh/

specificity is the antonym of generality, we only need to develop a computational model to estimate document generality. We hypothesize that two main factors, namely *document scope* and *document cohesion*, may influence the granularity of documents.

Document Scope (DS) reflects the topical coverage of a document. The lesser the number of domain-specific concepts present in the document, the larger the document scope will be. Moreover, with reference to the conceptual hierarchy of a domain ontology such as MeSH, the lower level domain concepts tend to have smaller document scope, whereas higher level domain concepts tend to have larger document scope. For example, a document containing the concept "Warts" is considered to have a smaller document scope (i.e., semantically more specific) than another document containing the concept "Virus Diseases". Figure 5 shows a fragment of the conceptual hierarchy encoded in MeSH. As can be seen, "Virus Diseases" is a higher level concept than "Warts" with reference to the MeSH conceptual hierarchy. Similarly, a document containing the concept "Condylomata Acuminata" is considered to have a smaller document scope than another document containing the concept "Warts".

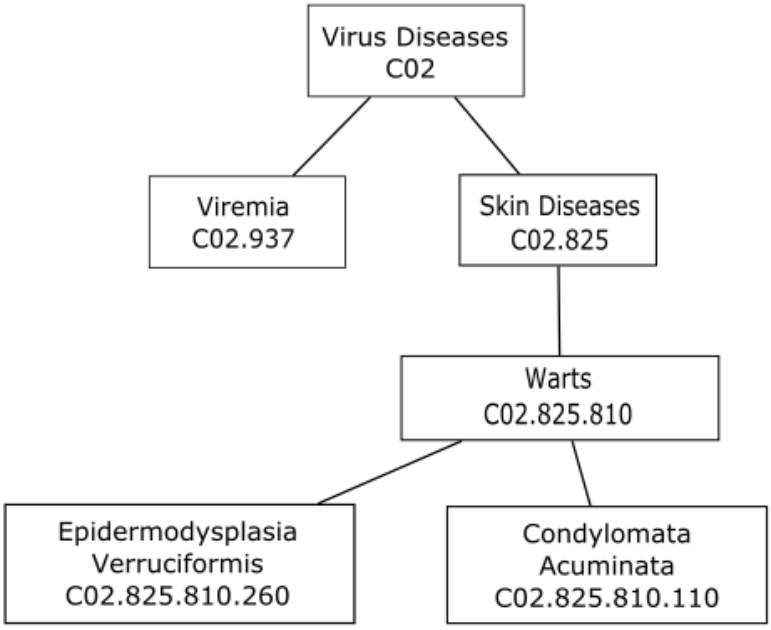

Fig. 5. An Example of Conceptual Hierarchy Encoded in MeSH

Document Cohesion (DC) reflects the semantic associations among the domain concepts appearing in a document. With reference to a domain ontology, the cohesion of a document can be measured in terms of the semantic association of the constituent terms. The more closely associated the terms are, the more cohesive the document tends to be. Derived from Ransdell's definition [Ransdell 1966; Santaella 2003], the semantic association of a text fragment refers to its capacity to represent a plurality of mutually

independent concepts with respect to a given domain. For example, given three short text fragments: T1 ("HIV1 and HIV2"), T2 ("HIV1"), and T3 ("HIV1 and Hypertension"), T2 is considered more cohesive than T1 as T2 only consists of a single concept. However, T1 is more cohesive than T3 because "HIV1" and "HIV2" are both sub-types of "HIV" and are less mutually independent than the concepts "HIV1" and "Hypertension" appearing in T3.

3.2 The Mesh Domain Ontology

Because document scope and document cohesion are estimated with reference to a given conceptual hierarchy encoded in a domain ontology, we provide an overview of the MeSH domain ontology used to illustrate and evaluate the proposed granular IR model. For the bio-medical domain, the controlled vocabularies, such as MeSH and SNOMED [7], that are part of the meta-thesaurus UMLS [8], provide a conceptual hierarchy where generalization relationships are defined among concepts called descriptors. Medical Subject Headings (MeSH) is used in our study because it is a controlled vocabulary for indexing MEDLINE, a popular online database containing 17 million medical and health related citations and abstracts. We treat document terms (including compounds) with matching counterparts in the MeSH domain ontology as MeSH identified concepts.

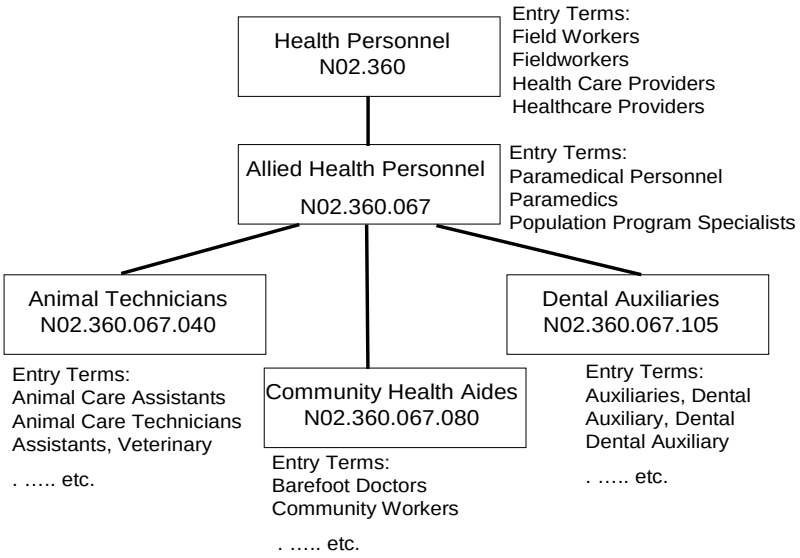

Fig. 6. A Fragment of MeSH Domain Ontology

Figure 6 is a fragment of the MeSH ontology where MeSH descriptors are numbered and organized based on a generalization relationship. For example, the heading "Allied

[7] http://www.snomed.org/
[8] http://umlsks.nlm.nih.gov/

Health Personnel" with a unique MeSH identification number N02.360.067 comes under "Health Personnel" (N02.360) , while "Community Health Aides" (N02.360.067.080) is listed under "Allied Health Personnel". Moreover, the MeSH ontology provides the "entry terms" which can be regarded as synonyms of a concept descriptor. In Figure 6, the heading "Allied Health Personnel" is linked to entry terms such as "Paramedical Personnel", "Population Program Specialists" and "Paramedics". The entry terms related to the MeSH concept descriptors facilitate the identification of MeSH concepts contained within bio-medical related documents.

3.3 Concept Identification

To estimate document scope and document cohesion, the domain concepts contained in documents need to be identified. For instance, both single and compound terms can be matched with the entry terms or the concept descriptors defined in a domain ontology such as MeSH. However, the situation becomes more complicated when a term (or several terms) matches more than one domain concept (e.g., a MeSH descriptor). For example, both the compound term "Plant Viruses" and the constituent term "Viruses" are MeSH descriptors. To effectively handle such a situation, we have developed the Conceptual Marking Tree (CMT) procedure based on a conceptual encoding technique [Zakos et al. 2003]. When compound words match more than one concept encoded in the ontology, the CMT algorithm can efficiently identify subsumed concepts presented at adjacent locations of a document. Then, the subsumed concepts will not be mistakenly identified as matching MeSH concepts. This procedure enables us to more accurately estimate document scope and document cohesion by identifying the correct domain concepts which appear in a document.

The CMT is a tree structure for storing the occurrence positions of all the document terms with matching domain concepts. A CMT is created and initialized for each document during the concept identification process. The structure of a CMT is similar to a domain ontology except that document offset arrays are provided in the CMT to record the locations of domain concepts found in a document. For the MeSH domain ontology, each node in the CMT corresponds to a MeSH concept descriptor. The semantics of the links between concept nodes in the CMT is the same as that of MeSH. A set of MeSH descriptors $C = \{c_1, c_2, \cdots, c_n\}$, such as "Plant Viruses", is represented as a concept node in the CMT. In addition, an *m*-word compound MeSH descriptor is described by a sequence such as $c_x = \langle c_{x1}, c_{x2}, \cdots, c_{xm} \rangle$ in the CMT. Each constituent concept term c_{xy} is associated with an array p_{xy} which is used to record the occurrence positions of that particular term in the document. An example of the CMT data structure is shown in Figure 7. The pseudo-code of the conceptual marking procedure *ConceptM* is provided in Table 1.

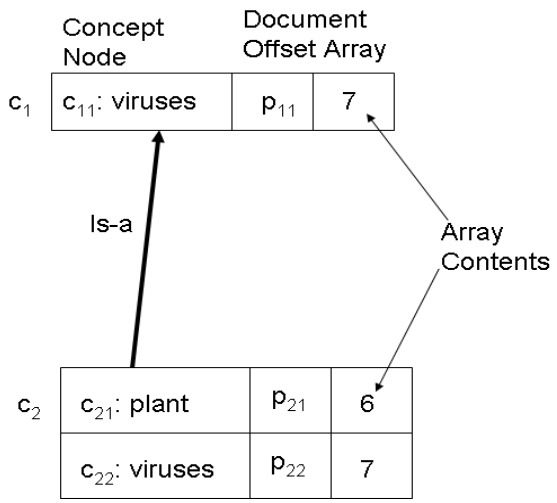

Fig. 7. A Fragment of the Conceptual Marking Tree

Table 1. The Procedure of Concepts Marking

Procedure *ConceptM*(D , Ont)
Inputs:
D /* a document
Ont /* a domain ontology
Output:
CMT /* a concept marking tree data structure
Main Procedure:
1. Initialize *CMT*; /* initialize CMT according to the Ont concept hierarchy
2. Pre-processing document *D*; /* stop word removal and stemming
3. Use sub-string matching to identify domain concepts appearing in *D*;
4. Record the document offset values of the matching concepts in CMT; /* concept marking
5. Check subsumed concepts in document *D*;
6. Reset the document offset values of the subsumed concepts in CMT;
7. Return *CMT*;

The main inputs to the conceptual marking tree algorithm include a document *D* and a domain ontology Ont, and the output of the algorithm is the conceptual marking tree *CMT*. There are basically seven steps to perform concept marking for a document. Step 1: a CMT data structure is created and initialized according to a particular domain ontology Ont. Step 2: a document *D* is parsed sequentially and traditional document pre-processing methods such as stop word removal and stemming are applied. Step 3: a standard sub-string matching function is applied to match domain concepts encoded in *CMT* with the tokens in the document *D*. Step 4: the locations of the matching domain

concepts in *D* are recorded in the *CMT*. Step 5: subsumed domain concepts are identified based on the document offset values stored in the *CMT*. Step 6: the document offset values are reset for the subsumed concepts. Step 7: the processed *CMT* data structure is returned to the caller. At the end of the concept marking procedure, the domain concepts that appear in the document will have non-zero document offset values recorded in the corresponding concept node of the *CMT*. For illustration, consider the following one-sentence document with respect to the MeSH domain ontology:

"Over 390 individual descriptions of plant viruses or plant groups are provided."

In this example, both "plant viruses" and "viruses" from the input document have matching concepts in the MeSH ontology after word stemming is performed. For readability reason, the original form of each word is shown in this example. The MeSH descriptor "plant viruses" has two elements, that is, c_{21}: "plant" and c_{22}: "viruses". As shown in Figure 7, the locations where these constituent concept terms appear in the document are recorded in the corresponding document offset arrays p_{21} and p_{22} respectively. For the MeSH descriptor "viruses", which has only one constituent word, it has one document offset array p_{11} in the CMT. As the document offset value of the concept "viruses" is the same as that of the constituent term "viruses" of the compound concept "plant viruses", "viruses" is taken as a subsumed concept and its corresponding document offset array will be reset. The end result is that only one domain concept "plant viruses" is identified in the aforementioned example.

Although it is possible to employ a simpler data structure and computational method for the concept identification process, additional computational time is still required to build a CMT-like data structure (i.e., a data structure similar to a domain ontology) for document cohesion computation at a later stage. The computational details about the estimation of document cohesion of a document are illustrated in Section 3.4.2 and Section 3.4.3 respectively. The advantage of using the conceptual marking tree (i.e., the data structure of a domain ontology) is that information about the relationships (e.g., semantic distances) between domain concepts is readily available for document cohesion calculation. There is no need to go through the process of marking the document terms against the domain ontology again for document cohesion computation.

We develop the CMT concept marking tool instead of using existing concept identification tools such as MetaMap [Aronson 2001; Bhatia et al. 2009] because we need a concept identification tool which is effective for a variety of application domains such as medicine, agriculture, computing, etc. We agree that MetaMap can be a viable alternative for concept identification for the medical domain. However, MetaMap is tightly coupled with the Unified Medical Language System (UMLS) and the underlying lexicon SPECIALIST™ which specializes in natural language processing for life science vocabulary. The proposed CMT method makes it easier for us to customize the concept

identification processes for different application domains. Our concept identification method was evaluated based on the OHSUMED-88 document subset; the system identified MeSH concepts were compared with the benchmark MeSH concepts which had been annotated as part of the OHSUMED collection. Our concept identification method achieved an average precision of 0.83. Such a result compares favorably with the published performance (in the range of 0.5 to 0.7) of MetaMap [Bhatia et al. 2009].

3.4 Re-Ranking Documents By Semantic Granularity

Given a query Q and a ranked list of documents $(R, \leq)$, where $R = \{d_1, \ldots, d_n\}$ is a set of documents, the objectives of our granular IR model will be:

- To construct a generality function $Gen(d_i)$ which returns the document generality $\forall d_i \in R$;
- To construct a new ranking function $f'(RScore(d_i, Q), Gen(d_i))$ which takes into account document similarity, popularity, and granularity;
- To re-rank documents by f' such that $f'(RScore(d_i, Q), Gen(d_i)) \leq f'(RScore(d_j, Q), Gen(d_j))$ implies $rank(d_i) \geq rank(d_j)$, where $d_i, d_j \in R$ and $rank(d_i)$ returns the rank of a document d_i.

Please note that a small rank number assigned to a document indicates that the corresponding document is placed at the top of a ranked list of documents. We hypothesize that the generality of a document, i.e., *Gen(d_i)*, could be estimated based on the scope and the cohesion of the document. $RScore(d_i, Q)$ refers to the combined similarity and popularity score computed by another component of our system, or imported from an external IR system.

3.4.1 Document Scope

The function $Scope(d_i)$ measures document scope by computing the average tree depths of all the domain concepts appearing in a document. The tree depth of a domain concept is measured by the distance between the concept node and the root node with reference to a particular concept hierarchy encoded in a domain ontology [Yan et al. 2006]. For any document terms that are not found in the domain ontology, their tree depths are defined to be zero. It may be the case that a document contains a large number of terms but only a few matching domain concepts. Consequently, the scope values of this kind of document may be skewed. To make the scope function sensitive to the average tree depths of the matching domain concepts, an exponential function is introduced to the document scope formula:

$$Scope(d_i) = e^{-\left(\frac{\sum_{i=1}^{n} depth(c_i)}{n}\right)} \quad (1)$$

In Equation 1, n refers to the total number of terms (i.e., matching domain concepts and non-matching terms) found in a document d_i. The function $depth(c_i)$ counts the tree depth of concept c_i with reference to a concept hierarchy. If a document contains only non-matching terms, its average tree depth is zero. Hence, it will have the maximum document scope of 1. For the MeSH domain ontology, the minimum document scope is e^{-11}, as the maximum tree depth of the MeSH hierarchy is 11. Therefore, for the MeSH domain, document scope values fall into the range [e^{-11}, 1]. As can be seen, the document scope function is a monotonically decreasing function with respect to the increasing average tree depth of the concepts contained in a document. The time complexity of scope-based document ranking is $O(MN)$, where M is the number of retrieved documents, and N is the total number of terms (i.e., both matching concepts and non-matching terms) contained in a document collection.

For terms not found in the domain ontology, their document scope may be estimated with reference to a generic lexicon such as WordNet. However, WordNet contains a large number of concepts that are not relevant to the target domain and these 'noisy' concepts may lead to an inaccurate estimation of concept specificity and concept cohesion. A previous study has compared the term relationships extracted from WordNet with those dynamically discovered using a text mining method for expanding the specific queries stored in a user profile [Lau et al. 2008]. Unfortunately, the WordNet based method led to poor IR performance. Therefore, to achieve a more accurate estimation of document scope and document cohesion, we focus on the concepts found in a domain ontology. The following examples demonstrate the computation of document scope.

Example One

Figure 8 shows two documents d_i and d_j with the constituent concepts mapped to a concept hierarchy. For ease of exposition, it is assumed that each document only contains two terms. The terms o and p in d_i and the terms k and h in d_j are the matching concepts found in the concept hierarchy (the darkened nodes). According to our document scope computation, the average tree depths of d_i and d_j are 3 and 4, respectively. Accordingly, the scope of d_i, 0.0498, is greater than the scope of d_j, 0.0183.

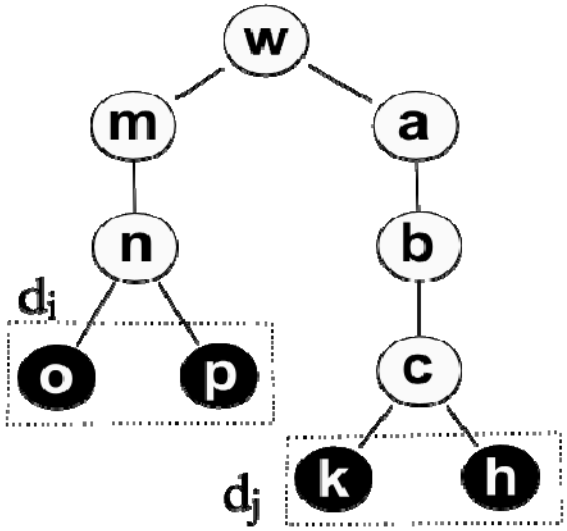

Fig. 8. D_i And D_j With Different Document Scope

Example Two

Figure 9 illustrates a document containing concepts that have subsumption relationships with each other, i.e., one concept is the parent node of another. The concepts m and n in d_i and c and h in d_j are the matching concepts found in the concept hierarchy (the darkened nodes). The average tree depths of d_i and d_j, respectively, are $(1+2)/2$ = 1.5 and $(3+4)/2$ = 3.5. The scope of d_i, 0.2231, is greater than the scope of d_j, 0.0302.

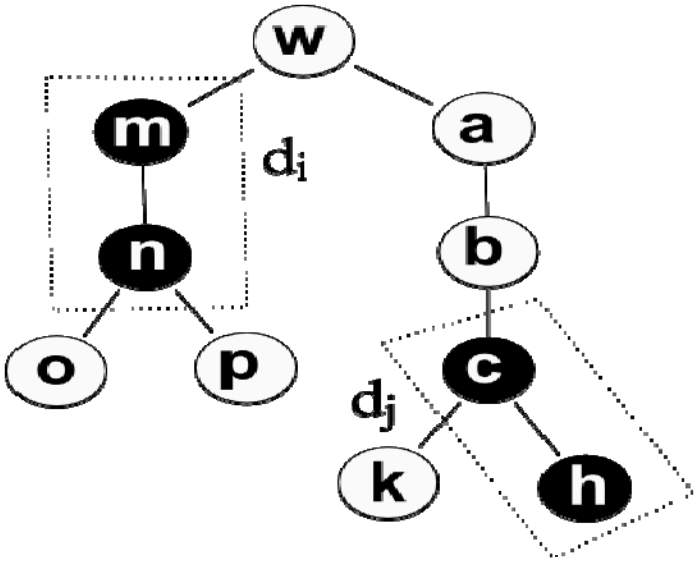

Fig. 9. d_i and d_j with subsumed concepts

3.4.2 Distance Based Document Cohesion

Document cohesion is a state or quality that the elements of a text "tend to hang together" [Morris and Hirst 1991]. The computation of document cohesion only considers concepts found in a domain ontology. If no concept is found in a document, the minimum cohesion value of zero is assumed. We adopt the Leacock-Chodorow similarity function [Leacock and Chodorow 1998] as one of the methods to estimate document cohesion, and refer to it as DC. The Leacock-Chodorow similarity function has been applied to measure the semantic similarity between two concepts by referring to a linguistic ontology [Leacock and Chodorow 1998]. The basic intuition of the Leacock-Chodorow function is that the semantic similarity between two concepts is estimated based on the conceptual links (i.e., the distance) between these concepts with reference to an ontology. The smaller the distance between two concepts, the higher will be the semantic similarity between these concepts. This approach also assumes that the links between two concepts represent uniform distances.

The reason why the Leacock-Chodorow similarity function is used to develop the proposed document cohesion function is that the Leacock-Chodorow similarity function was found more effective than an information content based similarity function according to a benchmark test involving 28 noun pairs [Resnik 1995]. When the cohesion value of a document is estimated, the average similarity of all pairs of matching concepts found in a document will be computed. The minimum similarity value is zero if the shortest distance between a pair of concepts equals to the maximum tree depth. The maximum similarity value is 1 if two matching concepts are directly linked to each other, or only one matching concept is found in a document. The document cohesion function is defined as follows:

$$Cohesion(d_x) = \frac{\sum_{(c_i, c_j \in MC) \wedge c_i \neq c_j} Sim(c_i, c_j)}{NumberofAssociations} \tag{2}$$

$$Sim(c_i, c_j) = -log \frac{len(c_i, c_j)}{2Depth_{MAX}} \tag{3}$$

$$NumberofAssociations = \frac{|MC|(|MC|-1)}{2} \tag{4}$$

where *MC* is the set of matching concepts found in a domain ontology. $Sim(c_i, c_j)$ is the Leacock-Chodorow semantic similarity function which takes into account the shortest path $len(c_i, c_j)$ between two concepts c_i and c_j defined in the domain ontology. $NumberofAssociations$ is the total number of associations among the set of matching concepts *MC*. $Depth_{MAX}$ is the maximal tree depth of a concept hierarchy. For the MeSH domain ontology, $Depth_{MAX}$ is 11. Accordingly, the document cohesion values fall in the range of $[0, -log(\frac{1}{22})]$. The time complexity of cohesion-based ranking is $O(m \times n^2)$, where m is the number of documents contained in a collection, and n is the number of concepts encoded in a particular domain ontology. The following examples illustrate the document cohesion computation.

Example One

Figure 10 illustrates the concept hierarchy of two documents d_i and d_j with their constituent concepts. The concepts o and p in d_i and x and y in d_j are found in the example hierarchy (the darkened nodes). The length of the shortest path between o and p is 2 and the shortest distance between x and y is 4. The document cohesion values of d_i and d_j are 2.3979 and 1.7047, respectively. Therefore, the contents of d_i are considered more cohesive than the contents of d_j.

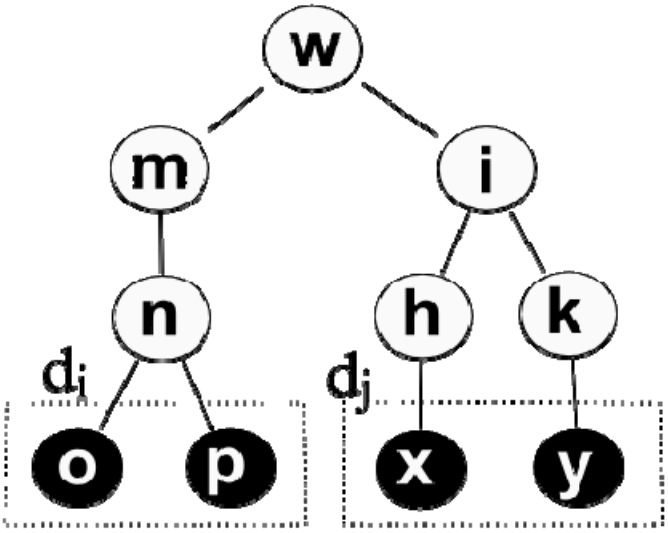

Fig. 10. d_i and d_j with different document cohesion

Example Two

Figure 11 illustrates a cohesion computation for documents containing subsumed concepts. The concepts o and n in d_i and i and y in d_j are found in the example hierarchy (the darkened nodes). The length of the shortest path between o and n is 1. The shortest distance between i and y is 2. Thus, the cohesion of d_i, 3.0910, is greater than the cohesion of d_j, 2.3979. Therefore, the contents of d_i are considered more cohesive than the contents of d_j.

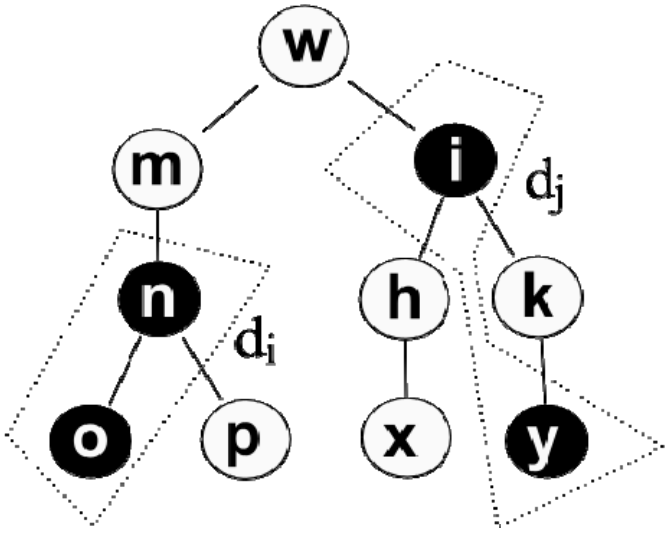

Figure 11. d_i and d_j with subsumed concepts

3.4.3 Information Content Based Document Cohesion

To alleviate the problem that the links in an ontology may not represent uniform distances, Resnik [Resnik 1995] proposed an information theoretic approach to measure the semantic similarity between concepts. For instance, two adjacent concepts at the top level may not have the same semantic similarity as another adjacent pair at the bottom level of the ontology. Resnik's method computes the information content of the least common subsumer (lcs) that subsumes the pair of target concepts. We consider this method to be an alternative to the distance-based approach to estimating document cohesion illustrated in Section 3.4.2, and refer to it as DC2. The semantic similarity of a pair of concepts is defined by [Resnik 1995]:

$$sim_{IC}(c_1, c_2) = \max_{c \in Subsum(c_1, c_2)} -\log_2 \Pr(c) \qquad (5)$$

$$\Pr(c) = \frac{\sum_{t \in Words(c)} tf(t)}{N} \tag{6}$$

$$Cohesion(d_x) = \frac{\sum_{(c_i, c_j \in MC) \wedge c_i \neq c_j} Sim_{IC}(c_i, c_j)}{NumberofAssociations} \tag{7}$$

where $sim_{IC}(c_1, c_2)$ represents the semantic similarity between concepts c_1 and c_2. The set $Subsum(c_1, c_2)$ represents the set of least common subsumers that subsume both c_1 and c_2, and $c \in Subsum(c_1, c_2)$ is one of the least common subsumers. $\Pr(c)$ represents the probability of a least common subsumer *c*, and is estimated based on the occurrence frequencies of terms subsumed by *c*. The set $Words(c)$ refers to *c* and all the terms subsumed by *c*. The term *N* is the total occurrence frequency of the constituent terms of the concepts observed in the reference corpus, and *tf(t)* is the occurrence frequency of the individual term *t*. For our experiment, we simply used the OHSUMED collection as our reference corpus in order to estimate *Pr(c)*. Equation 7 is similar to Equation 2 except that the similarity measure applied to a pair of concepts is based on the information contents of the least common subsumers rather than the absolute distance between the pair of concepts.

As an example, we use a fragment of the MeSH domain ontology depicted in Figure 5 to illustrate the computation of information content based semantic similarity. Based on the OHSUMED collection, the occurrences of the terms appearing in Figure 5 are as follows: (virus diseases, 5), (viremia, 109), (skin diseases, 171), (warts, 178), (epidermodysplasia verruciformis, 58), (condylomata acuminata, 106). Assuming that the concept hierarchy depicted in Figure 5 is a complete domain ontology, and the total term occurrence N = 5 + 109 + 171 + 178 + 58 + 106 = 627 is derived for our example collection, the semantic similarity between any pairs of concepts depicted in Figure 5 can be computed. For instance, the concepts "viremia" and "skin diseases" can be estimated based on their least common subsumer "virus diseases". The probability of "virus diseases" can be estimated according to Equation 6: $\Pr(\text{"virus diseases"}) = \frac{5+109+171+178+58+106}{627} = 1$. Accordingly, the semantic similarity $sim_{IC}(\text{"viremia","skin diseases"}) = -\log_2 1 = 0$ is estimated. Since we assume that the top node of this example ontology is "virus diseases", the value of its information content is zero. On the other hand, the semantic similarity between c_1 = "epidermodysplasia verruciformis" and c_2 = "condylomata acuminata" can be estimated based on their lcs "warts". The probability of "warts" is computed as follows: $\Pr(\text{"warts"}) = \frac{178+58+106}{627} = 0.545$, and the semantic similarity between the two concepts is $sim_{IC}(c_1, c_2) = -\log_2 0.545 = 0.874$.

As can be seen from the above example, the semantic similarity between "epidermodysplasia verruciformis" and "condylomata acuminata" is higher than that of

"viremia" and "skin diseases". Such a computational result matches our intuition about the similarities of these two pairs of concepts. As both "epidermodysplasia verruciformis" and "condylomata acuminata" are a kind of "warts", they are very similar diseases. However, "skin diseases" and "viremia", the disease caused by the presence of some viruses in the blood, are quite different diseases. Accordingly, a document containing the concepts "epidermodysplasia verruciformis" and "condylomata acuminata" is considered to have a higher document cohesion value than that of another document containing the concepts "skin diseases" and "viremia".

3.4.4 Overall Document Generality

Taking into account both document scope and document cohesion, we propose Equation 8 to estimate document granularity in terms of the overall document generality. We hypothesize that document generality *DG* is proportional to document scope and inversely proportional to document cohesion. Such a hypothesis will be evaluated based on both system-oriented experiments (Section 4.1) and user-oriented experiments (Section 4.2). The reason why we make such an assumption is that a document tends to be semantically general if its constituent terms are with large document scope (i.e., high level concepts with reference to a domain ontology). For example, a document containing the term "virus diseases" is considered more general than another document containing the term "warts" according to a domain ontology depicted in Figure 5. In addition, a document tends to be semantically general if its constituent terms are not cohesive. For example, a document containing the terms "skin diseases and viremia" is considered more general than another document containing the terms "epidermodysplasia verruciformis and condylomata acuminata". The reason is that "viremia" is a disease caused by the presence of some viruses in the blood and it is quite different from skin diseases. Accordingly, the corresponding document is semantically general because it covers different topics. On the other hand, both "epidermodysplasia verruciformis" and "condylomata acuminata" are specific kind of "warts", and so the corresponding document is semantically specific about the "warts" disease. With reference to the MeSH domain, the possible values of *DG* falls into the range $[\frac{e^{-11}}{-\log(\frac{1}{22})+1},1]$. The "+1" in the denominator of Equation 8 is to prevent the division by zero problem when the cohesion score of a document is zero. If the cohesion of a document is zero, document generality will be totally determined by document scope.

$$DG(d_i) = \frac{Scope(d_i)}{Cohesion(d_i)+1} \tag{8}$$

3.4.5 Absolute Document Re-ranking

Let $RScore(d_i, Q)$ denote the combined similarity and popularity score assigned to a document d_i. An aggregated document score $GRScore(d_i, Q)$ after taking into account semantic granularity is given in Equation 9.

$$GRScore(d_i, Q) = RScore(d_i, Q)^{\alpha} * e^{-\left(DG(d_i)^{\beta}\right)} \quad (9)$$

where α and β are the parameters for tuning the weights of the similarity/popularity component and the granularity component of the document ranking function respectively. The similarity/popularity score is assumed provided by another component of our IR system or provided by an external IR system such as an Internet search engine. For the extreme case where a document's *DG* score is close to zero (i.e., a document with high specificity), the granularity component of Equation 9 (i.e., $e^{-\left(DG(d_i)^{\beta}\right)}$) will approach one. Therefore, the aggregated document score $GRScore(d_i, Q)$ will be mostly determined by the similarity/popularity score. In other words, the most relevant and specific documents tend to be ranked at the top. For a document with a high *DG* score (i.e., general documents), its aggregated document score will decrease after applying Equation 9. The granularity component of Equation 9 is formulated as an exponential-decay function which can be considered as simulating Shepard's law [Shepard 1987], which states that exponential-decay functions are a universal law of stimulus generalization in psychology. In fact, a similar kind of exponential function has been successfully applied to estimate the semantic similarity of concepts [Li et al. 2003].

3.4.6 Document Re-ranking By Granularity Gap

To deliver the documents that best match an information seeker's specific needs, it also is important for an IR system to be able to estimate granularity requirements in terms of query generality. If we treat a query *Q* as a short document, Equation 10 can be used to estimate the query generality:

$$QG(Q) = \frac{Scope(Q)}{Cohesion(Q) + 1} \quad (10)$$

Moreover, Equation 11 is proposed to detect the granularity gap between a query and a document, and adjust the document rank automatically. The basic intuition is that if there is a large granularity gap between a query and a document (e.g., a specific query versus a general document), the initial similarity-based document rank should be adjusted (e.g., lowered). The reason is that the document is unlikely to meet the information seeker's granularity requirements. By incorporating the automatically estimated query generality into a document ranking function, it is possible to deal with variations of perceived

information granularity among different information seekers. For instance, a document one information seeker perceives to be specific, another may consider to be relatively general. Nevertheless, such variance will be reflected in the different terminologies employed in the respective queries.

The granularity gap between a query Q and a document d_i is estimated based on the absolute difference between $DG(d_i)$ and $QG(Q)$. For example, for a query with very low generality and a document with very high generality (i.e., $DG(d_i) > QG(Q)$), the granularity gap will be large and the granularity component of Equation 11 (i.e., $e^{-\left(|DG(d_i)-QG(Q)|^{\beta}\right)}$) will return a small number. Accordingly, the similarity/popularity score of the document will be multiplied by a small number. Consequently, the aggregated document score will become smaller, and the rank of the document is likely to be lowered.

$$GRScore(d_i, Q) = RScore(d_i, Q)^{\alpha} * e^{-\left(|DG(d_i)-QG(Q)|^{\beta}\right)} \quad (11)$$

4. EXPERIMENTS AND RESULTS

To verify the effectiveness of our proposed granular IR model, we applied a two-stage evaluation procedure. The first stage of the evaluation was a system-oriented experiment. We compared the IR effectiveness of the proposed granularity-based document ranking function with that of the traditional similarity-based ranking method in a bio-medical domain and an agricultural domain. For the second stage, we conducted user-oriented studies to compare our granularity-based document ranking function with the implicit ranking function exercised by information seekers. In addition, the user perceived relevance of the documents ranked by our granular IR system was compared with that produced by the Google search engine.

4.1 System-Oriented Experiments

4.1.1 The Document Collection

Our IR model was first evaluated based on the OHSUMED [Hersh et al. 1994] corpus, a subset of Medline containing 348,566 medical references. Each reference contains an abstract and a number of additional fields, such as title, author, source, and publication type. As with the TREC evaluation procedure [Hersh 1999], the abstract of each reference was regarded as a document in our experiment. The following is a fragment of a sample document from the OHSUMED collection, where:

I = Sequential Identifier
U = MEDLINE identifier
T = Title

P = Publication type
W = Abstract
M = Manually annotated MeSH concepts
A = Author
S = Source

.I 1
.U 87049087
.S
Am J Emerg Med 8703; 4(6):491-5
.M
Allied Health Personnel/*; Electric Countershock/*; Emergencies; ……..
.T
Refibrillation managed by EMT-Ds: incidence and outcome without paramedic back-up.
.P JOURNAL ARTICLE.
.W
Some patients converted from ventricular fibrillation to organized rhythms by defibrillation-trained ambulance technicians (EMT-Ds) will refibrillate before hospital arrival.
……
.A Stults KR; Brown DD.

4.1.2 Queries and Relevance Judgments

Apart from the document set, the OHSUMED collection consists of 106 topic descriptions and the corresponding relevance judgments [Hersh et al. 1994]. Each topic description contains two parts: the patient information (i.e., the title field) and the physician's information requirements (i.e., the description field). We used both the title field and the description field to construct a query, which is a standard method of constructing test queries for TREC-like experiments [Hersh 1999]. Of the 106 topics, 5 were dropped because they did not correspond to any relevant documents. As a result, a total number of 101 test queries were applied in our experiments.

4.1.3 Baseline Model and Document Pre-Processing

A well-known open source IR system, Lucene, was adapted to develop our baseline model for indexing and retrieving documents from the benchmark collections. Lucene is a popular benchmark system with basic keyword matching, inverted indexing and TF-IDF based term weighting functions [Salton and Buckley 1988]. All terms were filtered by the SMART 571 stop word list [Salton 1990] and then stemmed using the Porter stemmer [Porter 1980]. The same procedure was applied to our chosen domain

ontology. The TF-IDF weighting scheme was applied to index the test collections. For each domain, there are some rare cases that a domain concept descriptor or its related term also appears in the SMART stop word list if all characters are transformed to lower case format. For example, for the MeSH domain ontology, "LET" (Linear Energy Transfer), "UN" (United Nations), and "WHO" (World Health Organization) are concept descriptors and they might be mistakenly taken as stop words because the SMART stop word list contains "let", "un", and "who" as well. Our document pre-processing program employed a simple heuristic such that if a word comprised all upper case letters, it would be recognized as an acronym and its original case format would be retained. The same case transformation rule also was applied to process a domain ontology. As a result, MeSH concepts such as "LET" would not be removed from a document after our stop word removal process. On the other hand, ordinary English word such as "let" or "Let" would be removed because they were all converted to "let" and such a word was found in the SMART stop word list. For all the system-oriented experiments, the $RScore(d_i, Q)$ score as referred to in Equations 9 and 11 represent the similarity score only. The MeSH concepts appearing in each document were identified by using our conceptual marking tree algorithm.

4.1.4 Evaluation Methodology

In our experiments, the baseline IR system was used to retrieve 1,000 documents (ranked by similarity) for each test query. The granularity of each retrieved document was then computed and the 1,000 documents were re-ranked according to our combined similarity and granularity metric. We considered five granularity-based document ranking functions, in which document scope and document cohesion were used either alone or together to estimate document granularity.

The first experimental IR scenario used document cohesion alone to estimate document granularity. Then, a combined document similarity and granularity metric (i.e., Equation 9) was used to re-rank documents. In other words, documents with high similarity scores and low generality scores tended to be ranked at top positions. In this case, the overall document granularity is estimated based on solely document cohesion DC as shown in Equation 12. For the second experimental IR scenario, we adopted a similar experimental procedure except that the alternative cohesion function DC2, that is Equation 7, was deployed.

$$DG(d_i) = \frac{1}{Cohesion(d_i) + 1} \tag{12}$$

The third experimental IR scenario made use of only the document scope function DS to estimate document granularity. In other words, document granularity (measured in terms of overall document generality) was computed according to Equation 13. After computing document granularity, a combined document similarity and granularity metric

(i.e., Equation 9) was then applied to re-rank the documents as in the first experimental IR scenario.

$$DG(d_i) = Scope(d_i) \tag{13}$$

The fourth experimental IR scenario made use of the document scope and the document cohesion functions DS+DC to estimate document granularity. In other words, document granularity (measured in terms of overall document generality) was computed according to Equation 8. Documents were then re-ranked according to Equation 9 as with the previous IR scenarios.

The final experimental IR scenario also considered document scope and document cohesion for estimating document granularity. In addition, the document re-ranking mechanism tried to match documents of certain granularity level (measured in terms of document generality) with the corresponding granularity requirement induced from a query. We call this document ranking approach QS+QC+DS+DC. Basically, we treated each query as a short document and applied Equation 10 to estimate query granularity. Equation 11 instead of Equation 9 was then applied to re-rank the documents. Under such circumstances, documents having larger granularity gaps with respect to the query would be ranked closer to the bottom positions.

4.1.5 Evaluation Metrics

The IR performance of re-ranking documents using a combined similarity and granularity function was measured in terms of Mean Average Precision (MAP) and R-precision. Precision is the proportion of documents retrieved that are relevant. Recall is the proportion of the relevant documents that are retrieved. The MAP is computed across different recall points, i.e., when a relevant document is retrieved, and averaged over all the topics. R-precision represents the precision value after *R* relevant documents have been retrieved. These metrics are widely used to measure the performance of IR systems [Hersh et al. 1994].

4.1.6 Experimental Results

The precision-recall graph shown in Figure 12 compares the IR effectiveness of the baseline IR system (i.e., Lucene) and the proposed granular IR system. To highlight the improvement achieved by various granularity-based document ranking functions, Figure 13 presents a magnified view of a certain range of precision and recall levels in Figure 12. As shown in Figure 13, the DS ranking function outperforms all the other ranking functions in terms of MAP from recall level 0 to 0.4. Table 2 summarizes the performance improvement or deterioration due to the DS ranking function. The precision values as achieved by various ranking functions with respect to different recall levels are tabulated in Table 3. The MAP, R-Precision, the percentages of improvement (Δ%), and

the α and the β parameter values adopted under various granular IR scenarios are summarized at the bottom of Table 3. The symbols *** and ** appearing in the Δ% rows of Table 3 indicate statistically significant improvement at the levels of $p \leq .01$ and $p < .05$, respectively.

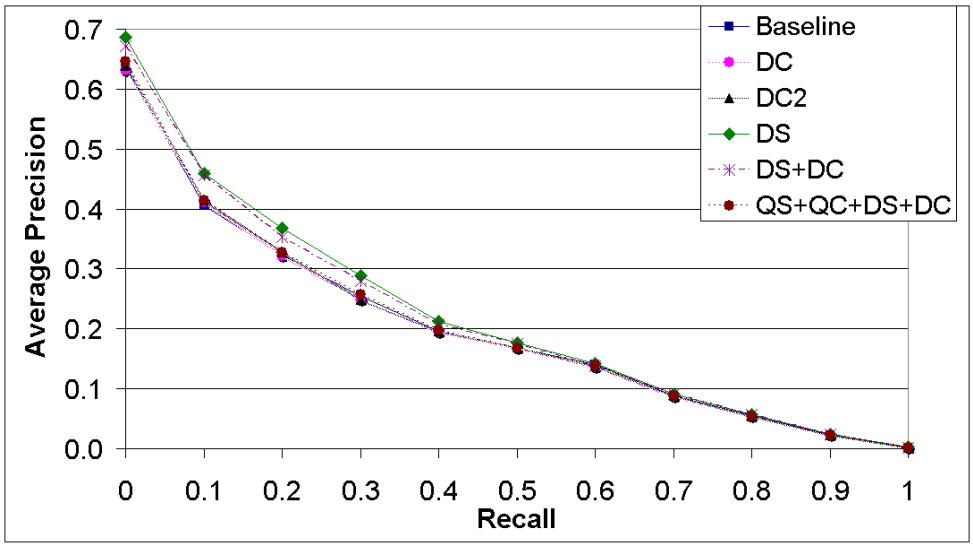

Fig. 12. Precision-Recall Graph for Overall IR Performance

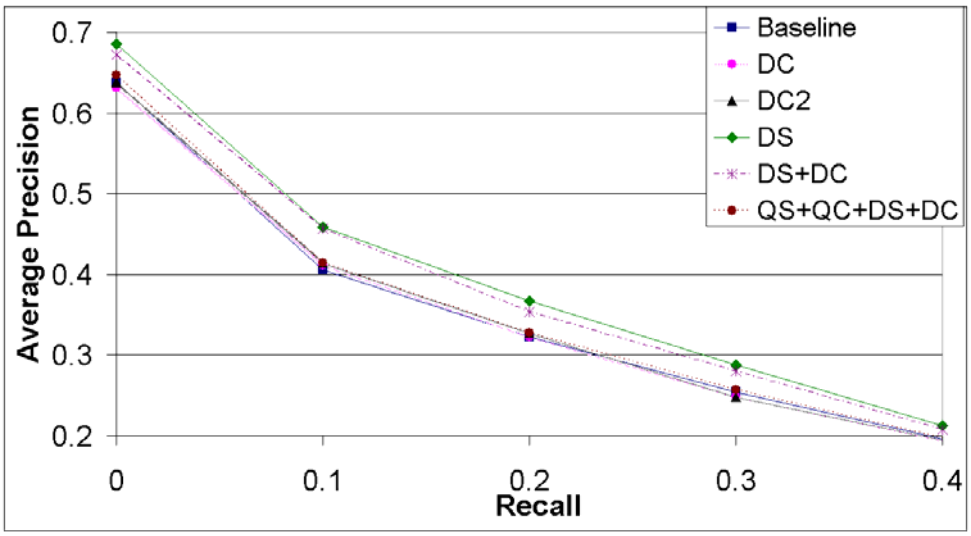

Fig. 13. Segment of Precision-Recall Graph for
Recall in [0, 0.4] and Precision in [0.2, 0.7]

Table 2. The Average Improvement/Deterioration of the DS Ranking Function

Total Number of Queries	101
Number of Queries with MAP improved	62
Average Improvement	23.76%
Number of Queries with MAP deteriorated	39
Average Deterioration	13.59%

Table 3. Detailed Precision-Recall Comparisons for the OHSUMED Collection

Recall	Baseline	DC	DC2	DS	DS+DC	QS+QC+DS+DC
0	0.6369	0.6311	0.6386	0.6858	0.6721	0.6465
0.1	0.4071	0.4110	0.4141	0.4591	0.4563	0.4124
0.2	0.3239	0.3222	0.3278	0.3674	0.3532	0.3276
0.3	0.2540	0.2480	0.2485	0.2881	0.2799	0.2536
0.4	0.1963	0.1942	0.1956	0.2125	0.2079	0.1973
0.5	0.1679	0.1681	0.1687	0.1770	0.1747	0.1680
0.6	0.1396	0.1364	0.1375	0.1414	0.1374	0.1390
0.7	0.0880	0.0873	0.0875	0.0917	0.0907	0.0879
0.8	0.0544	0.0537	0.0545	0.0565	0.0562	0.0543
0.9	0.0223	0.0221	0.0223	0.0236	0.0234	0.0222
1	0.0018	0.0018	0.0018	0.0023	0.0021	0.0018
MAP	0.1849	0.1834	0.1851	0.2036	0.1994	0.1865
Δ%		-0.81%	0.11%	***10.11%	**7.84%	1.30%
R-Prec	0.2246	0.2234	0.2247	0.2800	0.2446	0.2268
Δ%		-0.53%	0.04%	***24.67%	**8.90%	0.98%
(α, β)		(6, 1)	(6, 1)	(4, 1)	(5, 1)	(2, 1)

The values of the α and the β parameters were tuned according to our empirical testing conducted based on the OHSUMED collection. To have a more fine-grained tuning, we estimate the parameter values with respect to each granularity-based document ranking function employed in our system. We tried different combinations of the α and the β values, and observed the corresponding MAP results to find a good combination (e.g., a local optima). The results of parameter tuning for the various granularity-based ranking functions such as DS, DC, DS+DC, QS+QC+DS+DC are plotted in Figure 14. The value of β is always smaller than the value of α according to our parameter tuning process. However, such a result matches our intuition about document relevance. The content of a document should be similar to (about) a query in the first place. Then, the proposed granularity-based document ranking function is applied to fine-tune the result of a similarity-based ranking such that it also satisfies a user's granularity requirement.

Therefore, the α parameter which controls the similarity part should be assigned a heavy weight.

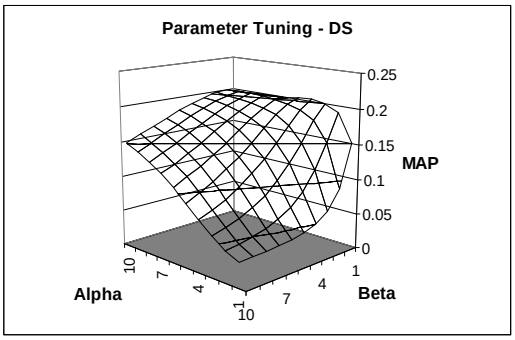

Fig. 14a. Parameter Tuning for DS Function

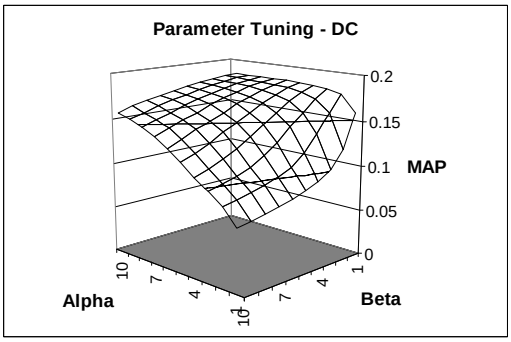

Fig. 14b. Parameter Tuning for DC Function

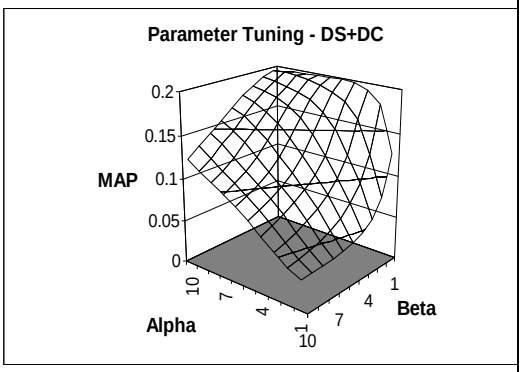

Fig. 14c. Parameter Tuning for DS+DC Function

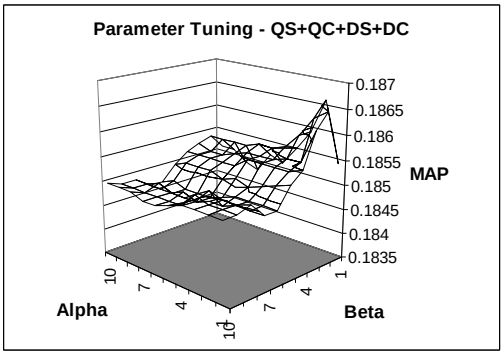

Fig. 14d. Parameter Tuning for QS+QC+DS+DC Function

Fig. 14. Parameter Tuning for Granularity-Based Document Ranking

After tuning the system parameters in one training domain, we then apply them to other application domains. This is a kind of cross-domain tuning and validation. In particular, we applied the tuned parameters to perform document ranking in the agricultural domain (Section 4.1.8) and the IT domain (Section 4.2.4, Section 4.2.5, Section 4.2.6), respectively. If the performance of the proposed document ranking functions is good in these domains, it will demonstrate that these parameters have been properly tuned. Another alternative is to apply an in-domain parameter tuning approach where a subset of queries is used to search for reasonable parameters, and then these parameters are applied to the rest of the queries in the same domain. However, there is a trade-off between tuning effort and system performance. By using the cross-domain parameter tuning approach, we aim at minimizing the human effort in tuning the parameters for each test domain. Nevertheless, it is interesting to see if in-domain parameter tuning can further bootstrap the performance of our granular IR system in future research.

4.1.7 Result Analysis

Among all the variations of granularity-based document ranking, the DS ranking function achieved the best IR effectiveness. By applying the DS ranking function, performance improvement in terms of MAP was achieved in 62 out of the 101 test queries. The overall average improvement was 10.11% in terms of MAP and 24.67% in terms of R-Precision for all the test queries. We performed a paired one tail *t*-test which compared the paired mean precision values achieved by the baseline IR system and the DS granularity-based IR system over all the test queries. The *t*-test result showed that statistically significant difference was found ($t(10) = 3.07$, $p = .01$). Therefore, we conclude that the improvement brought by the DS ranking function is statistically significant. Moreover, by applying the DS+DC ranking function, significant improvement of MAP also was achieved ($t(10) = 2.81$, $p = .02$). However, the degree of improvement was not as large as that brought about by the DS ranking function. Both the QS+QC+DS+DC and the DC2 ranking functions only achieved marginal improvement in terms of MAP and R-Precision, and these results are not statistically significant. These initial experimental results show that some of our granularity-based ranking functions such as DS and DS+DC are effective complements to the traditional similarity-based ranking functions.

According to these experiments, it appears that the IR effectiveness can be significantly improved by applying the Document Scopes (DS) based ranking function. Given a document with contents bearing certain similarity with the query, the smaller the document scope (i.e., very specific about a topic), the more likely it is judged as relevant. Most of our test queries were specific queries in a bio-medical domain. Therefore, documents with specific bio-medical contents rather than general information were more likely to satisfy the specific bio-medical queries. These experimental results match our basic intuition about domain-specific IR in the OHSUMED setting. In order to carry out a deeper analysis, we manually verify the ranked list of documents generated according to the DS ranking function. We use a test topic (39) and documents (317458 and 115871) of the OHSUMED collection to give an intuitive explanation of the improvement brought by the DS ranking function. The document titles are listed below with the matching MeSH concepts highlighted in italic.

Topic #39 – "35 Y O WITH ***GASTROENTERITIS VIRAL GASTROENTERITIS***, CURRENT MANAGEMENT"

DID	Document
317458	***Calicivirus gastroenteritis*** in a long-term care facility for the elderly.
115871	***Viral gastroenteritis***. As our ability to control many of the common ***infectious diseases*** has increased, attention has turned toward the less common or less severe ***infections***. It is clear that worldwide, significant numbers of the cases of ***gastroenteritis*** in both adults and children are caused by ***viruses***. Many of these ***viruses*** now are quite well understood and their control appears to be on the horizon. Many other ***etiologic agents*** are just being identified and will present a challenge to researchers and practitioners alike.

For the test topic (39), the MAP increases from 21.43% (i.e., baseline) to 26.15% by performing the DS based re-ranking. A non-relevant document (317458) is ranked 21st by the baseline system. However, another relevant document (115871) is ranked even lower than document (317458) at the 28th rank. It can be observed that the title of document (317458) contains fewer number of MeSH concepts than document (115871). After DS based re-ranking (i.e., applying Equation 9), document (317458) is ranked lower (e.g., at the 28th rank) because of its higher document generality. On the other hand, the rank of document (115871) is boosted up to 12th because of its higher document specificity. The overall IR effectiveness is improved because of the above document re-ranking process.

Table 2 shows that the proposed granularity-based document ranking functions such as DS can improve IR performance for many topics, but it also may degrade performance for some topics. The reason is that the DS ranking function only blindly decreases the document scores for general documents with reference to the MeSH ontology. However, general documents (e.g., documents containing only a few MeSH concepts with small tree depths) also can be relevant with respect to a relatively general query such as OHSUMED Topic 4 "reviews on subdurals in elderly". As a result, performance degradation may occur in such an occasion. Nevertheless, both the DS and the DS+DC document ranking functions demonstrate statistically significant improvement of MAP over the baseline system. In fact, only partial functionality of the proposed system can be examined in a controlled system-based experiment. Therefore, we apply a series of usability studies to supplement the system-oriented experiments. The series of usability studies will be described in Section 4.2.

On the other hand, the DC ranking function led to a slight degradation of IR performance. The reason is that general documents tend to have high level concepts found in the MeSH domain ontology. Nevertheless, the absolute distances among these high level concepts

tend to be small as well because they are located close to the root node of the concept hierarchy. According to Equation 2, the cohesion of such documents tends to be high, and the generality of these documents becomes relatively low. In other words, these documents could be mistakenly treated as specific, and they were ranked at higher positions than they should have by our document ranking function depicted in Equation 9. Unfortunately, these general documents do not really match the specific OHSUMED queries adopted in our experiment. As a result, the overall IR performance was degraded. Intuitively, employing an information content based metric [Resnik 1995] could avoid such a problem. However, significant improvement was not found by employing Resnik's semantic similarity measure. The reason may be that the term distribution statistics of the particular reference corpus (i.e., OHSUMED) we adopted in the experiment do not match well with the semantic granularity of the concepts as encoded in the MeSH ontology. As a result, the DC2 ranking function did not contribute much in improving the performance of the IR tasks. Indeed, another study on semantic similarity of concepts also found that the information content based method did not resemble well how humans perceive the similarity of semantically related concepts [Li et al. 2003].

It is surprising to find that the QS+QC+DS+DC ranking function, which tries to re-rank documents according to the degree of match between the document granularity and the implicit granularity requirement embedded in a query, does not lead to the greatest improvement in performance. According to our experiments, the QS+QC+DS+DC ranking function only brings a marginal MAP improvement of 1.3%. A further investigation into the OHSUMED collection reveals that most of the test queries are long and containing many MeSH concepts, such as the names of symptoms and illness. In other words, most of these queries are specific in nature. In addition, the OHSUMED corpus is a domain-specific collection (e.g., it only contains medical references). The granularity gap between a test query and an arbitrary document is small in general. Therefore, little adjustment may arise after applying Equation 11 to re-rank the documents. For instance, with a zero granularity gap, the similarity score of a document will not be modified by Equation 11 at all. In contrast, the IR performance significantly improves after the DS ranking function is applied because it forces a re-ranking by promoting the rank of the specific documents, which is what is really expected from the specific OHSUMED queries. Another reason why little improvement is brought by the QS+QC+DS+DC ranking function may be that the proposed DC formulation is not effective as shown by our experimental results. As a result, the overall effectiveness of the QS+QC+DS+DC ranking function is deteriorated.

As granularity gap is not easily quantified, future research is required to refine the current computational method to accurately estimate granularity gaps between documents and queries. For instance, a specific document about "role for tumor necrosis factor-alpha in

JC virus reactivation and progressive multifocal leukoencephalopathy" retrieved from PubMed may exhibit a large granularity gap with respect to the query "an overview about tumor". Having a robust computational method to estimate the granularity gap between the document and the query is crucial under such a circumstance. Since the granularity gap between the document and the query is large in this case, the document should be ranked lower toward the bottom of the ranked list when compared to other general documents about "tumor". Otherwise, too specific and probably irrelevant information will be delivered to the user. On the other hand, for the query "swine diseases" (C22.905), the document about "Porcine Reproductive and Respiratory Syndrome" (C22.905.700) should be ranked higher toward the top of the ranked list than another document about "diseases" (C) because the granularity gap between "swine diseases" and "Porcine Reproductive and Respiratory Syndrome" is smaller than that between "swine diseases" and "diseases" according to the MeSH ontology. In fact, "Porcine Reproductive and Respiratory Syndrome" is an instance of "swine diseases". There may be occasions that evaluating the granularity gap between a user query and a document cannot improve IR effectiveness at all. One such an example is when the similarity between a query and a document is low. For instance, both the query "warts" (C04.925.744) and the document "systems integration" (L01.906.787) have the same semantic granularity (i.e., document scope and document cohesion) with reference to the MeSH ontology. However, re-ranking the document according to the match of their semantic granularity cannot improve IR effectiveness since the content of the document bears little similarity with that of the query.

It should be noted that the best MAP achieved in the 2006 TREC Genomics Track was slightly higher than 0.5 [Hersh et al. 2006]. However, the MAP value obtained in our experiment should not be directly compared to that announced in the 2006 TREC Genomics Track because of different experimental environments. For example, there were 162,259 full-text documents used in the 2006 Genomics Track, while the OHSUMED collection used in our experiment contained 348,566 medical abstracts only. Essentially, our baseline system was implemented based on the well-known vector space model which has been widely used in IR research [Salton et al. 1975; Salton 1990; Salton 1991]. Having such a commonly used baseline system makes it easier to compare our experimental results with other published results. The proposed document ranking function leads to statistically significant improvement of MAP over the baseline method. However, since the MAP achieved by the baseline system is not very good, more experiments which involve stronger baseline methods such as Okapi BM25 [Robertson et al. 1998] should be conducted to further evaluate the merits of the proposed document ranking function in the future.

4.1.8 Cross Domain Evaluation

To examine whether our proposed granular IR model is effective in another domain-specific IR scenario, we conducted a system-oriented experiment based on the Reuters-21578 collection [9] and the FAO's AGROVOC [10] domain ontology for agricultural topics. The AGROVOC contains 28,954 concepts (descriptors) and some other related terms, and the depth of this ontology is 10, which is close to that of the MeSH ontology. Figure 15 shows a segment of the AGROVOC ontology which includes the concept "sugar" and its related concepts. In Figure 15, the number below each concept descriptor is the unique concept identification number of the AGROVOC ontology.

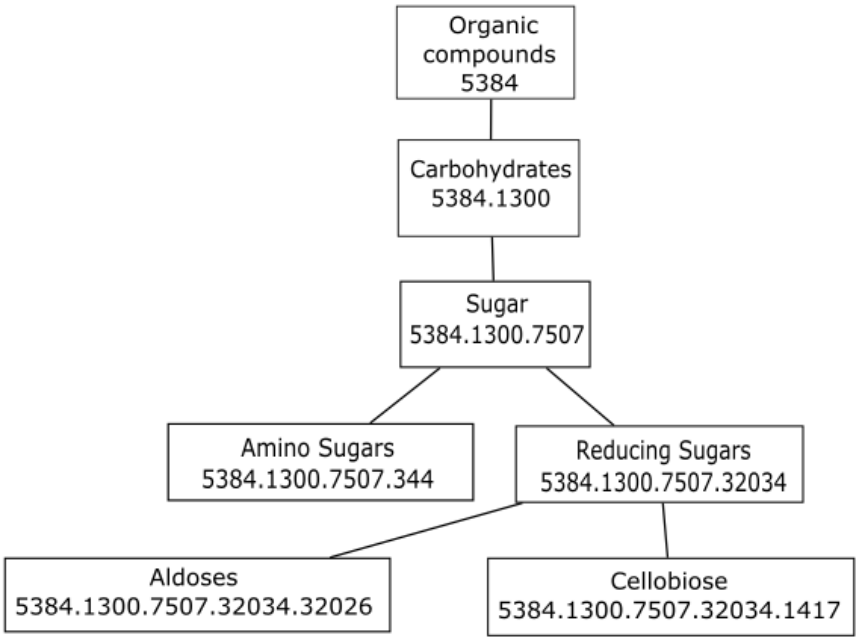

Fig. 15. A Segment of the AGROVOC Ontology

In this experiment, we used the Reuters-21578 collection together with the Lewis-Split subset containing 19,813 documents. A document parsing procedure was applied to generate the TREC like relevance judgment file[11] and the respective queries. For example, if the topic code "sugar" is found in the <Topics> field of a Reuters-21578 document, a relevance judgment record will be created. The following is a sample document with the embedded topic code "sugar" from the Reuters-21578 collection:

[9] http://www.daviddlewis.com/resources/testcollections/
[10] http://aims.fao.org/en/website/AGROVOC-Concept-Server/sub
[11] http://trec.nist.gov/data/web/09/prels.1-50.gz

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5589" NEWID="46">
<DATE>26-FEB-1987 15:51:28.42</DATE>
<TOPICS><D>sugar</D></TOPICS>
<TEXT>
<TITLE>U.S. SUGAR IMPORTS DOWN IN WEEK - USDA</TITLE>
<BODY>Sugar imports subject to the U.S. sugar import quota during
the week ended January 9, the initial week of the 1987 sugar quota
year, totaled 5,988 short tons versus 46,254 tons the previous week,
the Agriculture Department said. The sugar import quota for the 1987
quota year (January-December) has been set . . . . .
</BODY>
</TEXT>
</REUTERS>
```

Each Reuters-21578 document is delimited by the pair of <reuters> and </reuters> tags. Within the <reuters> tag, there are attributes to describe the unique document number and other header information. The pair of <body> and </body> tags are used to describe the main textual contents of the document. The above example only shows a segment of the whole article. The other tags are used to define the title of the Reuters news article and the date of publication. After document pre-processing, the topic field of each Reuters-21578 document was removed. The Reuters-21578 topic file (part of the Reuters-21578 collection) containing 135 pre-defined topics was used to construct the initial queries. To form a query, the topic code was used to extract the corresponding terms from the Reuters-21578 category description file. For instance, the corresponding terms of the corporate code "gnp" is "gross national domestic product". Usually, a topic description consists of one or two terms only. For the above topic code "sugar", only one term is available to describe the topic. Twenty topics such as "sun-oil", "coffee", "soybean", "sugar", etc. with relevant documents and matching concepts found in the AGROVOC ontology were randomly chosen to evaluate the performance of the granularity-based document ranking function. In general, the test topics of the Reuters-21578 collection are semantically more general than those of the OHSUMED collection used in our first experiment.

The experimental procedure and the values of the system parameters were the same as those applied to the OHSUMED-based experiment. The overall experimental results are reported in Table 4, and the topic-by-topic MAP and R-Precision scores of the DS ranking function are shown in Table 5. The symbols ** and * appearing in the Δ% rows of Table 4 indicate statistically significant improvement at the levels of $p < .05$ and p

$< .10$, respectively. As a whole, both the MAP and the R-Precision values were lower than that of the OHSUMED-based experiment because the domain-specific IR tasks for this agricultural domain were more difficult. Basically, only short queries with one or two terms were applied in these IR tasks. The DC, DC2, and QS+QC+DS+DC ranking functions marginally improve the MAP. However, these improvements are not statistically significant. On the other hand, both the DS and the DS+DC ranking functions performed significantly better than the baseline ranking function does. By means of paired one tail *t*-test, it was confirmed that the improvements brought by the DS and the DS+DC ranking functions were statistically significant with ($t(10) = 2.61$, $p = .03$) and ($t(10) = 1.98$, $p = .06$), respectively. As a whole, the results of our system-oriented experiments demonstrate that both the DS and the DS+DC ranking functions can significantly improve IR effectiveness across different domains. As the set of system parameters estimated based on the medical domain also can lead to promising results in the agricultural domain, it suggests that the set of empirically established system parameters is properly tuned.

Table 4. Detailed Precision-Recall Comparisons for the Reuters-21578 Collection

Recall	Baseline	DC	DC2	DS	DS+DC	QS+QC+DS+DC
0	0.5536	0.5536	0.5537	0.5549	0.5542	0.5537
0.1	0.3194	0.3196	0.3197	0.3238	0.3215	0.3202
0.2	0.2321	0.2323	0.2325	0.2353	0.2341	0.2332
0.3	0.1605	0.1608	0.1612	0.1659	0.1647	0.1624
0.4	0.1022	0.1023	0.1025	0.1094	0.1069	0.1038
0.5	0.0665	0.0670	0.0680	0.0803	0.0765	0.0710
0.6	0.0456	0.0458	0.0462	0.0580	0.0540	0.0478
0.7	0.0292	0.0295	0.0298	0.0380	0.0344	0.0312
0.8	0.0115	0.0117	0.0119	0.0192	0.0160	0.0129
0.9	0.0016	0.0016	0.0018	0.0069	0.0047	0.0025
1	0.0003	0.0003	0.0003	0.0006	0.0004	0.0003
MAP	0.1157	0.1159	0.1162	0.1226	0.1201	0.1174
Δ%		0.19%	0.46%	****5.99%**	*3.86%	1.49%
R-Prec	0.1470	0.1470	0.1472	0.1507	0.1493	0.1478
Δ%		0.00%	0.15%	*2.56%	1.57%	0.57%

It should be noted that the published best IR performance for the Reuters-21578 collection was in the range of [0.80, 0.92] in terms of accuracy [Dumais et al. 1998; Liu et al. 2002; Peng et al. 2004]. However, previous research employed supervised classifiers such as a Naive Bayes classifier with hundreds of labeled documents as training examples [Pang et al. 2004]. For a supervised classification task, some relevant and some non-relevant documents are made available to a classifier (e.g., a Naive Bayes

classifier) during a training phase. The classifier then uses the labeled documents to learn the discriminatory features (e.g., terms) to classify a document. During the test phase, the classifier makes use of learned features to classify unlabeled test documents as relevant or not. Nevertheless, for our document ranking task (i.e., a TREC routing task), only a query is available to our document ranking mechanism; the document ranking mechanism must rank the collection of documents by placing the most relevant documents at the top of the ranked list immediately; training examples (documents) are not available to the document ranking mechanism to learn the prominent features to distinguish between relevant documents and non-relevant documents. Accordingly, the document ranking tasks reported in our experiments are more difficult than the supervised classification tasks reported in previous studies [Dumais et al. 1998; Peng et al. 2004].

Table 5. Topic-by-Topic Performance of the DS Ranking Function

Topic	No. Relevant Documents	MAP	R-Prec
grain	305	0.1472	0.3856
wheat	251	0.1286	0.2351
sorghum	137	0.1075	0.1195
lin-oil	25	0.0339	0.0092
sun-oil	120	0.1025	0.1088
soybean	192	0.5988	0.7043
earn	2445	0.2317	0.4298
housing	182	0.1129	0.2298
coffee	296	0.1034	0.1109
ship	183	0.0666	0.0105
sugar	548	0.1027	0.1091
carcass	114	0.0988	0.0695
livestock	625	0.0689	0.0603
crude	130	0.0872	0.0675
nat-gas	112	0.0922	0.0679
cpi	163	0.0832	0.0651
copra-cake	43	0.0713	0.0499
plywood	62	0.0693	0.0406
wool	216	0.1002	0.1001
heat	64	0.0459	0.0399

Furthermore, topic-by-topic performance optimization also was conducted to improve the effectiveness of document classification in previous research [Dumais et al. 1998]. For instance, the priori probability distribution *Pr(c)* of a topic *c* (i.e., a class) was established

based on labeled training documents for each individual topic [Dumais et al. 1998]. In contrast, our proposed document ranking mechanism did not make use of any labeled training documents to learn the classification features or probability distributions of terms from any query topics. As the same document ranking procedure was applied to every query topic in our experiments, topic-by-topic IR optimization was not performed. Nevertheless, both the DS and the DS+DC document ranking functions perform significantly better than the baseline system in terms of MAP. Therefore, we can conclude that document scope is an important granularity factor which can improve IR effectiveness. As our current baseline system only produces a relatively low MAP, stronger baseline systems should be tried to examine if the same percentage of performance improvement can be achieved by our proposed document ranking functions in future experiments.

Post-experiment analysis found that both the DS and the DS+DC ranking functions could improve the average precision by promoting the positions of some relevant documents toward the top of the ranked list for semantically specific topics such as "sorghum". By applying DS or DS+DC based documents re-ranking through Equation 9, semantically general documents were assigned smaller document scores. In other words, semantically general documents were ranked toward the bottom of the ranked list, and at the same time the positions of the semantically specific documents were promoted toward the top of the ranked list. For example, the Reuters-21578 document (15500) with title "RPT - Argentine Grain/Oilseed Export Prices Adjusted" contains many semantically specific concepts such as "sorghum" with tree depth 4, "linseed oil" with tree depth 4, "soybean oil" with tree depth 4, "pollard" with tree depth 5, etc. This document is relevant with respect to the query "sorghum". The document was ranked 13th by the baseline system. However, after applying the DC document ranking function, it was promoted to the 6th position of the ranked list. As a result, the average precision was improved.

However, as there are relatively more semantically general topics in the Reuters-21578 collection than are in the OHSUMED collection, the percentage of performance improvement brought by the DS and the DS+DC document ranking functions is smaller when compared to that achieved by applying the same ranking functions to the OHSUMED collection. For example, for the general topic "housing", quite a number of relevant documents are semantically general with reference to the AGROVOC ontology. One such example is document (3105) with document title "Canada Building Permits Rise in November" which describes the changes of the number of building permits in Canada. By applying the DS document ranking function, the position of this relevant document was moved from the 15th place to the 16th place. Obviously, the DS document ranking function did not help improve the average precision in this case. A similar document re-ranking result was observed for the DS+DC document ranking function. As

a result, the overall percentage of performance improvement brought by the DS or the DS+DC document ranking function was smaller when they were applied to the Reuters-21578 collection. On the other hand, the DC ranking function might move the semantically general documents up in a ranked list due to the same reason explained in Section 4.1.7. Nevertheless, placing general documents at higher positions of a ranked list is likely to produce a better match with the general queries. Therefore, there is a slight performance improvement brought by the DC document ranking function in this experiment.

Since only some combinations of the granularity factors have been examined by our system-oriented experiments, evaluation for other granularity factors or their combinations (e.g., QS+DS) should be conducted in the future. The current approach only considers in-document cohesion or in-query cohesion; another alternative is to consider a combined query-document cohesion measure to rank documents with respect to their semantic relatedness to the query. In addition, other baseline systems can be considered. One such candidate is to use a cosine-similarity function to directly measure the overlap of domain concepts between a query and a document. From a theoretical perspective, our proposed approach differs from such a baseline method in that not only the overlapping concepts are considered for document ranking but also the levels of specificity (or generality) of the matching concepts are evaluated. The proposed granular IR model can take into account the particular specificity or generality requirement imposed in a user query.

4.2 User-Oriented Evaluation

System-oriented experiments help evaluating the effectiveness of particular document ranking functions. However, these experiments cannot examine the full operational characteristics of the proposed granular IR system. A series of user-oriented studies were conducted to supplement the system-oriented experiments. Since only the DS and the DS+DC granularity-based document ranking functions show statistically significant improvement over the baseline system, we focus on these two factors, namely DS and DS+DC, in the remaining user-based studies. In particular, we examined the following research questions:

- How well do our formulations of the document scope and the document cohesion functions approximate the human perception of document scope and document cohesion for certain information items?
- How well does the DS+DC document ranking function approximate the implicit document ranking function exercised by humans?
- Does our granular IR system perform better than another IR system employing a combined similarity and popularity based document ranking function?

4.2.1 Evaluation Methodology

The problem with user-oriented IR experiments is that large benchmark IR collections, such as the OHSUMED corpus, may impose an excessive cognitive load on the human subjects. Therefore, we developed small collections of short document passages to examine the human implicit ranking function. The passages were developed with respect to a medical domain and an IT domain. MeSH was used as the medical domain ontology and the 1998 ACM Computing Classification System[12] was selected as the IT domain ontology. As in the system-oriented experiments, we examined both the document scope factor DS and the document cohesion factor DC. The combined effect of document scope and document cohesion DS+DC was then examined. Finally, a usability study was conducted to compare the perceived IR effectiveness of our system with that of Google. The aim of the DS, DC, and DS+DC ranking tests is to examine how well our specific formulations approximate the corresponding functions exercised by humans. The purpose of the perceived IR effectiveness test is to evaluate if our proposed granularity-based document ranking function, particularly the DS+DC function, can improve IR effectiveness or not.

The DS, DC, and DS+DC ranking tests were performed in the medical domain (26 participants with a major in nursing) and in the IT domain (32 participants with a major in IT). The usability study of our prototype system involved 48 undergraduate students with a major in nursing. All the participants were randomly chosen for our experiments and each participant voluntarily answered a set of questions regarding the ranking of selected document passages or the perceived effectiveness of an IR system. The participants were not introduced to the concepts of document scope and document cohesion during the experiments. A sample of the questionnaire used for this study can be downloaded from our project Website[13].

4.2.2 Document Sources

To allow the participants and our granular IR system to rank documents with various granularity, we employed a general search engine (e.g., Google) to retrieve some passages with high document scope. In addition, we utilized a domain-specific search facility provided by PubMed to retrieve passages with low document scope from the medical domain. Each passage was then presented to the participants to elicit their perception of its granularity. The documents for the ranking tests conducted in the IT domain mainly came from Google and domain-specific digital libraries, such as the ACM digital library.

[12] http://www.acm.org/about/class/1998
[13] http://quantum.is.cityu.edu.hk/GranularUserTest2009.doc

4.2.3 Evaluation Metrics

To compare the rankings of documents produced by human subjects with that generated by our system, we employed both the simple matching coefficient and the Spearman rank-order correlation coefficient [Gan et al. 2007]. The Spearman rank-order correlation coefficient is a widely used correlation analysis method for ordinal data. With reference to the confusion matrix depicted in Table 6, the simple matching coefficient is derived from:

Table 6. A Confusion Matrix for Comparing System/Human Judgment

Human's Judgment	System's Judgment		
		1	0
	1	a	b
	0	c	d

$$sm = \frac{a+d}{a+b+c+d} \tag{14}$$

where *sm* is the simple matching coefficient, and *a, b, c*, and *d* refer to the number of observations falling into each category. In addition, the Spearman rank-order correlation coefficient r_s is defined by:

$$r_s = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2\right)}{n(n^2-1)} \tag{15}$$

where *n* is the number of ranks for comparison, and d_i is the difference between two corresponding ranks.

4.2.4 Ranking Tests for Document Scope

In general, it may be difficult for information seekers to produce a fine-grained ranking of documents solely based on document scope. Therefore, we developed a comparative ranking method to elicit a subject's ranking of documents with respect to document scope. For instance, a pair of snippets referring to the description of a particular disease was presented to a subject. Each subject then decided (ranked) which snippet would be more specific than another one according to their own perception. When we constructed the ranking tests, the description of a disease only differed in terms of the number MeSH concepts used and the tree depth of these MeSH concepts. For example, technical terms were used in one of the descriptions of a disease, and layman words also were applied to describe the same disease. The following pair of descriptions is an example of our test for

document scope:

A. Avian influenza is a contagious disease caused by influenza A viruses found chiefly in birds, but infections also can occur in humans. (low document scope)

B. Avian influenza is caused by viruses found chiefly in birds, but it also may affect humans. (high document scope)

Both snippets have more or less the same document cohesion because they refer to the same disease. For this experiment, the definitions of ten diseases such as avian influenza, malignant neoplasms, Alzheimer's disease, Parkinson disease, fibrosarcoma, etc. were used to elicit the subjects' rankings. Our granular IR system also would use the DS document ranking method to rank each pair of snippets. Then, a human's ranking and our system's ranking of snippets with respect to document scope can be compared based on the simple matching coefficient. A high simple matching coefficient indicates that our system's ranking function DS can closely approximate a human's implicit ranking function based on document scope. The set of system parameters used in the system-oriented experiments was applied to the series of usability studies. The experimental results are shown in Table 7 below:

Table 7. User-Oriented Document Scope Test

Medical Domain	*sm*	IT Domain	*sm*
Avian Influenza	1.0	Java	1.0
Malignant Neoplasms	1.0	SQL	0.9375
Alzheimer's Disease	0.8523	Power Macintosh	1.0
Parkinson Disease	1.0	SCSI	1.0
Fibrosarcoma	0.9231	CMU SLM Toolkit	0.8125
Asthma	0.9231	AutoCAD	1.0
Botulism	1.0	Oracle	1.0
Stroke	1.0	Graph Theory	0.7813
Anaphylaxis	0.8523	DOS	1.0
Autoimmune Disease	1.0	IBM	1.0
Overall *sm*	0.9551	Overall *sm*	0.9531

Each participant's judgment was compared with the results produced by our granularity-based IR system DS according to the simple matching coefficient metric defined by Equation 14. A mean score was then computed for each granularity test which was related to the description of a disease. The first two columns of Table 7 show the results of the ranking tests in the medical domain, and the remaining two columns show

the results obtained in the IT domain. As shown at the last row of Table 7, the average simple matching coefficients for the medical domain and the IT domain are 0.9551 and 0.9531 respectively. The inter-rater agreement of the document rankings as measured by the Kappa value [Fleiss 1971] for the IT domain (0.8545) is very close to that obtained from the medical domain (0.8501). The similar Kappa value reflects that the difficulty of evaluating the document scope of medical terminologies, or the document scope of IT terminologies is more or less the same for human subjects who possess adequate domain knowledge. The high average SM values in both domains reveals that the proposed DS ranking function closely approximates humans' implicit document ranking functions.

4.2.5 Ranking Tests for Document Cohesion

Similarly, the document cohesion tests also required participants to compare a pair of passages. In particular, the participants were asked to judge which passage was the more cohesive with respect to the symptoms of a particular disease or the different descriptions about an IT topic, such as C# programming. To avoid the comparison of document scope between the pair of passages, both passages contained the same number of concepts with respect to the particular domain.[14] The following is an example of our document cohesion test:

A. C# is a programming language which has properties similar to that of SQL and Java. (low document cohesion)
B. C# is a programming language which has properties similar to that of C++ and Java. (high document cohesion)

Table 8 presents the results of the document cohesion tests in the same manner as the results of the document scope tests depicted in Table 7. The result of our test reveals that human subjects do refer to document cohesion to distinguish the granularity of two documents when their document scope is more or less the same. The average simple matching coefficients are 0.7311 for the medical domain and 0.7469 for the IT domain, respectively. The proposed DC document ranking function demonstrates a certain degree of correlation to the ranking functions exercised by human subjects. However, since the SM values obtained from this test are lower than those observed in the document scope test, there is room for improvement regarding the current formulation of our DC function. In fact, the inter-rater agreement as measured in terms of Kappa is only 0.6583 for the IT domain and 0.6291 for the medical domain, respectively. This result suggests that it may not be an easy task to determine the cohesion of two apparently similar snippets even for human subjects who have appropriate domain knowledge. Such a result partly explains why both the DC and the DC2 ranking functions could not achieve good IR performance

[14] As the sums of tree depths of the domain concepts appearing in two passages may not be exactly the same, the influence of document scope is controlled but it is not totally eliminated in this study.

in the system-oriented experiments. The implication is that, rather than considering document cohesion alone, the proposed granularity-based document ranking function should take into account multiple factors. Refinement of the document cohesion based document ranking functions will be left as part of our future work.

Table 8. User-Oriented Document Cohesion Test

Medical Domain	*sm*	IT Domain	*sm*
White Blood Cells	0.7292	C#	0.1250
Blood Cells	0.6769	Vector Graphic Editors	0.0310
Blood Diseases	0.7292	Computer Game	0.9688
Molecular Basis of Neuromuscular Diseases	0.6769	Database Management Systems	0.9375
Treatment of Lung Diseases	0.7876	Data Structure	0.8750
Impact of Animal Diseases	0.7292	Microprocessors Structure	0.8438
Autoimmune Basis of Connective Tissue Diseases	0.7876	Integrated Programming Environments	0.9375
Skin Diseases	0.7292	Text-based Operating Systems	0.8750
Risk of Negative Emotions	0.7876	Word Processing Tools	0.9375
Pain Effect of Using Magnifying Tools	0.6769	Web Applications	0.9375
Overall *sm*	0.7311	Overall *sm*	0.7469

4.2.6 Combined Document Ranking Tests

In each combined document ranking test, subjects were asked to rank four snippets. In particular, the first two snippets vary in terms of document scope (e.g., discovered by relating the terms in the snippets to the concepts defined in a domain ontology), and the remaining two snippets differ in terms of document cohesion. For each domain, five tests were developed to elicit the subjects' rankings. The subjects were told to rank the snippets in terms of their specificity in the medical domain or the IT domain. Our granularity-based IR system also was invoked to rank the same set of snippets using the DS+DC document ranking function. The strength of the DS+DC document ranking function was not shown to the participants. The following is an example of our combined

granularity test:

A. Tree is a structure that emulates a hierarchical data structure with a set of linked nodes. (medium DS and high DC)
B. Tree is a structure applied to computing. (high DS and low DC)
C. Object-oriented programming (OOP) is a programming paradigm that uses "objects" and their interactions to design applications and computer programs.
(low DS and high DC)
D. Tree is a hierarchical data structure and Object-oriented programming (OOP) is a programming paradigm. (medium DS and low DC)

As shown in Table 9 and Table 10, there are strong positive correlations between the humans' rankings and that produced by our system's ranking function DS+DC for both the medical domain ($r_s = 0.8215$) and the IT domain ($r_s = 0.8262$). The Spearman confidence levels are listed in the third column of each table; these values can be found in the Spearman critical values table[15]. If the confidence level is less than or equal to 5%, it is sufficient to reject the null hypothesis and conclude that the correlation relationship exist; otherwise the relationship is not statistically significant. As shown in Table 9 and in Table 10, the confidence levels of all our tests are not greater than 1%, and so the correlations between the human subjects' document rankings and our system's document ranking are statistically significant.

Table 9. Combined Ranking Test (Medical)

Medical Tests	Spearman Correlation	Confidence Level
Gastroesophageal Reflux	0.8231	≤1%
Malignant Neoplasms	0.8308	≤1%
Alzheimer's Disease	0.8308	≤1%
Parkinson Disease	0.8154	≤1%
Fibrosarcoma	0.8077	≤1%
Overall	0.8215	

Table 10. Combined Ranking Test (IT)

IT Tests	Spearman Correlation	Confidence Level
Tree Structure	0.8250	≤1%
CMU SLM Toolkit	0.8250	≤1%
AutoCAD	0.8250	≤1%
Oracle	0.8375	≤1%
Graph Theory	0.8188	≤1%
Overall	0.8262	

[15] http://psychology.ucdavis.edu/SommerB/sommerdemo/correlation/hand/critvalues_rs.htm

As a whole, the results of the user-oriented document ranking tests confirm that the proposed DS function is an effective approximation of how humans perceive the granularity of information items. In contrast, the proposed DC function is not as effective as the DS function for approximating the human perception of document granularity. As for the DS+DC function, it demonstrates a high correlation with the human perception of document granularity, probably because of the dominating influence of the DS part in the overall DS+DC function. The results of these user-oriented studies provide a cognitive justification of why the DS and the DS+DC document ranking functions improved retrieval effectiveness in the system-based experiments.

4.2.7 Usability Studies of The Granular IR System

To gain greater insight into how information seekers perceive the effectiveness of our granular IR system, we conducted two usability experiments. For the first experiment, 22 participants used our granular IR system and the Google search engine to conduct 5 search tasks in the medical domain. The purpose of the first experiment was to directly compare the users' perceived effectiveness of the two IR systems. For the second experiment, 26 participants who had not taken part in the first experiment were recruited to employ two different document ranking functions supported by our granular IR system to conduct another five domain-specific search tasks. The main purpose of the second experiment was to evaluate users' perceived effectiveness of both our proposed document ranking function and the Google document ranking function under the same user interface. Although the first usability study directly compares the participants' perceived effectiveness of the two systems, the perceived difference of IR effectiveness may be a result of the different interface design rather than the underlying document ranking functions. Our second experiment further examines if any difference in the participants' perceived IR effectiveness is really caused by the respective document ranking functions or not. All the participants involved in the usability studies had basic knowledge of the medical domain.

Our granular IR system employed the Google Search API[16] to retrieve the top 50 Web documents from Google, and then re-ranked these documents in descending order of specificity. Since the Google similarity/popularity score $RScore(d,Q)$ of a returned document d was not available via the API, our system employed a monotonically decreasing function $RScore(d,Q)=1-\left(\frac{rank(d)-1}{N}\right)$ to generate the raw document scores such that the original Google ranking was preserved; the function $rank(d)$ returns the rank of a document d and $N=50$ is the total number of documents returned from the Google search engine. For the granular IR system, manual specification of the retrieval granularity was used. In particular, the participants were told to specify a high level of information specificity for each IR task by using the granularity control bar

[16] http://code.google.com/apis/ajaxsearch/

(like the slider bar of Google Maps) provided by our system. In other words, they were to move the granularity control bar to the right-hand side as shown in Figure 3. The DS+DC function was used for the granularity-based document ranking in these experiments since it could lead to statistically significant improvement of IR performance according to our system-oriented experiments. Moreover, the DS+DC function was a close approximation of the implicit document ranking function exercised by humans according to our combined document ranking test. In fact, the DS function also was used in these experiments and it achieved more or less the same results. For brevity, we only report the results of our usability studies based on the DS+DC function.

Table 11. A Sample of the Online Questionnaire for the Usability Study

Question 1 **Stiff-Person Syndrome** 1. Enter the query "causes sps" to the Granular IR system to retrieve information regarding the possible causes of Stiff-Person Syndrome; 2. Slide the Granularity Bar on the top of the search box to "Specific" i.e., Right Most; 3. Click the button "Search with Granularity" to conduct the Web search; 4. From the "Result Window", click the TOP 10 URLs to view the Web page information; 5. From your perspective, do you agree that the Web pages contain RELEVANT information regarding the causes of Stiff-Person Syndrome? [] 1. Strongly Disagree [] 2. Disagree [] 3. Neutral [] 4. Agree [] 5. Strongly Agree
Question 2 **Stiff-Person Syndrome** 1. Enter the query "causes sps" to the Granular IR system to retrieve information regarding the possible causes of Stiff-Person Syndrome; 2. Slide the Granularity Bar on the top of the search box to "Specific" i.e., Right Most; 3. Click the button "Search" to conduct the Web search; 4. From the "Result Window", click the TOP 10 URLs to view the Web page information; 5. From your perspective, do you agree that the Web pages contain RELEVANT information regarding the causes of Stiff-Person Syndrome? [] 1. Strongly Disagree [] 2. Disagree [] 3. Neutral [] 4. Agree [] 5. Strongly Agree

For each experiment, the participants were first informed of the objectives of the experiment and shown the basic operations of our granular IR system. Then, each participant followed our pre-written instructions to carry out each IR task. A maximum of 5 minutes were allowed to carry out each IR task. The order of the IR tasks assigned to each subject was randomized in each experiment. Similarly, the order of the particular IR system used in experiment one and the order of the particular document ranking function employed in experiment two also was randomized to avoid the side effects of sequential system (or ranking function) exposure. Each domain-specific IR task was developed using the query template "causes of <disease name>". Table 11 shows a sample of the prescribed instructions of the second usability study. After finishing an IR task, the

participant was told to indicate the perceived relevance of the top ten information items (i.e., Web pages) returned by the system according to a 5-point semantic differential scale of "Strongly Agree (5)", "Agree (4)", "Neutral (3)", "Disagree (2)", to "Strongly Disagree (1)".

For the first experiment, the independent variables are IR systems and IR tasks, and the dependent variable is users' perceived relevance of the search results. It is basically a 2 (systems) by 5 (IR tasks) factorial design. In fact, similar kind of usability study was applied to examine semantic component based domain-specific search before [Price et al. 2007]. We used diseases such as African Horse Sickness (AHS), Postpartum Depression (PPD), Attention Deficit Disorder (ADD), Pelvic Inflammatory Disease (PID), and SARS to construct the corresponding IR tasks. The results of the first experiment are summarized in Table 12. The mean and the standard deviation (STD) of the users' perceived relevance scores are tabulated under the "Mean" and the "STD" columns respectively. The two-way ANOVA indicates no significant interaction between IR systems and IR tasks, $F(4, 210) = .37, p = .83$, partial $\eta^2 = .01$, but significant main effect for IR systems, $F(1, 210) = 18.74, p < .01$, partial $\eta^2 = .08$. The mean scores of perceived relevance of the granular IR system are consistently higher than that of the Google search engine for all the IR tasks. Therefore, we conclude that the perceived relevance of the results produced by our granular IR system is significantly higher than that produced by the Google search engine.

Table 12. The Perceived Relevance Tests of Different Systems

	Granular IR System		Google	
	Mean	STD	Mean	STD
Task1: causes of AHS	4.136	0.710	3.591	0.734
Task2: causes of PPD	4.182	0.733	3.818	0.501
Task3: causes of ADD	4.318	0.716	4.045	0.785
Task4: causes of PID	4.227	0.612	3.909	0.684
Task5: causes of SARS	4.091	0.750	3.545	0.739
Overall	4.191	0.697	3.782	0.709

Further investigation of this experiment revealed that a combined similarity and popularity ranking function, as employed by Google, may not always produce the most relevant and specific results. As shown in Figure 2, the first document returned by Google in response to the AHS query was a Facebook community page about a charity network of AHS, rather than the medical causes of AHS. On the other hand, our granular IR system returned more relevant and specific documents about the possible causes of AHS by evaluating the semantics carried by the documents. Figure 4 shows the top documents returned by our granular IR system with respect to the same AHS query. Another example is related to the query about SARS. Our granular IR system returned all the medical pages related to SARS in the top 10 list. On the other hand, the news page about a traffic jam caused by "South African Revenue Service" officials on strike was

ranked fifth by Google in September 2009. As a result, the participants in our study consistently perceived the search results produced by our granular IR system to be more relevant.

For the second experiment, the independent variables are document ranking functions and IR tasks and the dependent variable is the participants' perceived relevance of the search results. In this experiment, we compared the IR effectiveness of our DS+DC document ranking function with the Google document ranking function based on the same user interface. The participants' were told to conduct an IR task by clicking the "Search with Granularity" button, which invoked the DS+DC document re-ranking, or by clicking the "Search" button, which did not invoke any document re-ranking function. In the latter case, this meant that our granular IR system displayed the search results using the exact ranking produced by the Google search engine. A sample of the participants' prescribed instructions is depicted in Table 11, and a snapshot view of the system interface is shown in Figure 3. It should be noted that the order of invoking the DS+DC or Google document ranking function and the IR tasks performed were randomized for each participant. In this experiment, illnesses such as Acquired Brain Injury (ABI), Traumatic Brain Injury (TBI), Orofacial Myofunctional Disorders (OMD), Lambert-Eaton Myasthenic Syndrome (LEMS), and Stiff-Person Syndrome (SPS) were used to develop the corresponding IR tasks.

Table 13. The Perceived Relevance Tests of Different Ranking Functions

	DS+DC Ranking		Google Ranking	
	Mean	STD	Mean	STD
Task1: causes of ABI	4.231	0.504	4.077	0.549
Task2: causes of TBI	4.269	0.592	4.154	0.533
Task3: causes of OMD	4.346	0.676	3.846	0.769
Task4: causes of LEMS	4.385	0.487	4.231	0.421
Task5: causes of SPS	4.346	0.476	3.731	0.710
Overall	4.315	0.547	4.008	0.597

The results of the second experiment are summarized in Table 13. The two-way ANOVA indicates no significant interaction between document ranking functions and IR tasks, $F(4, 250) = 1.99$, $p = .09$, partial $\eta^2 = .03$, and no significant main effect for IR tasks, $F(4, 250) = 1.60$, $p = .18$, partial $\eta^2 = .03$. However, there is a significant main effect for document ranking functions, $F(1, 250) = 17.48$, $p < .01$, partial $\eta^2 = .07$. The mean scores of perceived relevance of the results generated by the DS+DC document ranking function are consistently higher than that of the Google document ranking function for all the IR tasks. Therefore, we can conclude that the perceived relevance of the results produced by the DS+DC document ranking function is significantly higher than that produced by the Google ranking function. The reason why the DS+DC document ranking function can outperform the Google ranking function is that information seekers'

granularity requirements are taken into account by the DS+DC ranking function. Figure 16 and Figure 17 show the search results due to the DS+DC ranking and the original Google ranking following a query concerning the causes of SPS. As the figures illustrate, the top results produced by the DS+DC ranking function are all related to the Stiff-Person Syndrome. However, the second, third, and forth entries among Google's top 10 ranking have no relation to the Stiff-Person Syndrome. The results of this experiment also reveal that information specificity does play an important role in determining the overall perceived relevance of information.

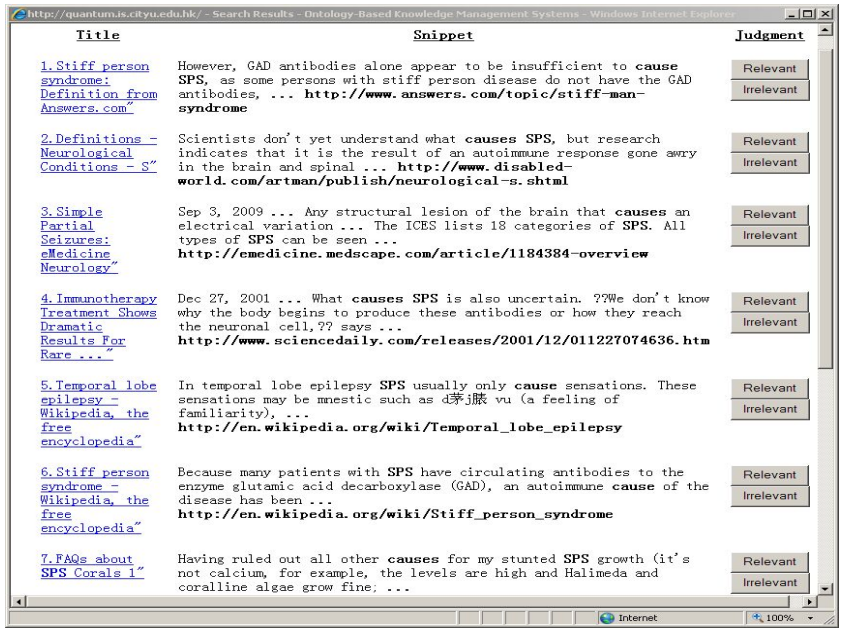

Fig. 16. Search Results Generated by DS+DC Ranking

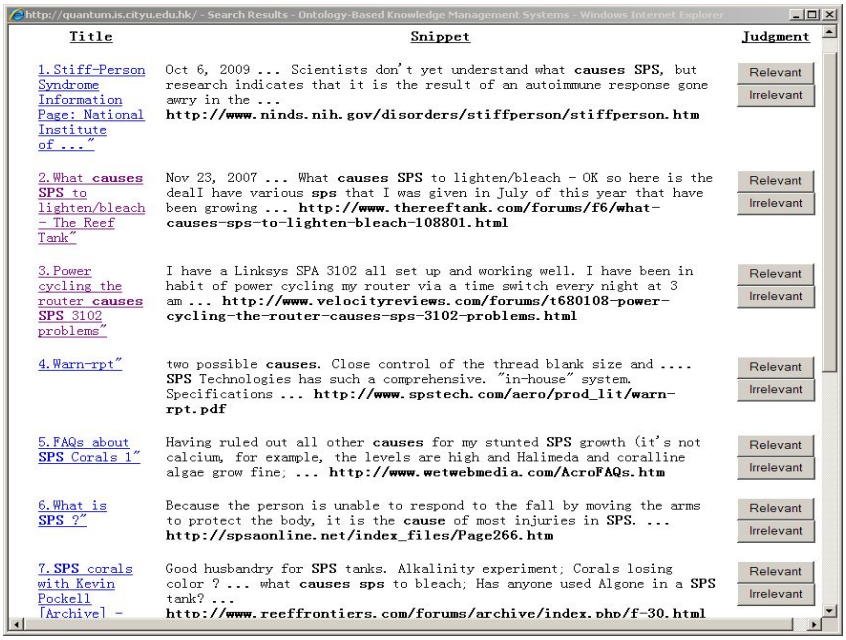

Fig. 17. Search Results Based on the Original Google Ranking

5. CONCLUSIONS AND FUTURE WORK

Although similarity-based and popularity-based document ranking functions have been successfully applied to IR in general, other factors should be considered to improve the effectiveness of domain-specific IR. In this paper, we propose a novel granular IR model which takes into account three dimensions, namely "similarity", "popularity", and "granularity" to enhance domain-specific IR. In particular, the notions of document scope and document cohesion are introduced to estimate the semantic granularity of documents or queries. A novel computational method for granularity-based document ranking is developed. System-oriented experiments, based on benchmark document collections and domain ontology pertaining to a medical domain and an agricultural domain, were conducted to evaluate the effectiveness of the proposed method. Our experimental results reveal that IR effectiveness significantly improves when the document scope function is applied to measure the granularity of documents and re-rank these documents accordingly. Series of user-based ranking tests also show that granularity-based document ranking based on the document scope function or the combination of document scope and document cohesion function closely resembles the implicit human ranking function. In addition, information seekers perceived our granular IR system to be able to deliver more relevant results than the Google search engine for some domain-specific IR tasks.

Future research should examine a more effective document cohesion based ranking function and to incorporate such a function into our proposed granularity-based document ranking mechanism. For instance, not only document or query cohesion would be measured individually but also the combined query-document cohesion should be considered for document ranking. The effectiveness of other combinations of the granularity factors also should be tested. Moreover, the computational method for estimating the granularity gap between a query and a document needs to be refined and empirically tested with a larger scale usability study. More sophisticated parameter tuning methods (e.g., genetic algorithms) will be applied to search for optimal or near optimal system parameters. Another line of research is to apply the ontology extraction method [Lau et al. 2009a] to automatically build a domain ontology, so that the proposed granularity-based ranking function can be applied to arbitrary information domains. The proposed granular IR model can be extended to support general IR scenarios where documents may not contain domain-specific concepts. Under such circumstances, statistical term frequency computed based on a collection of documents or a general lexicon, such as WordNet, may be applied to estimate the document scope of terms.

APPENDIX - Correlation Between Granularity and Similarity

Figures A1 to A4 show the correlation analysis between document similarity (x axis) and document granularity (y axis). A dot in a diagram indicates the corresponding similarity score and granularity score of a retrieved document. In particular, the similarity score of a document is computed using our baseline IR system (Lucene). The granularity score of the document is computed according to Equation 8 (i.e., the DS+DC method). Because of the limitation of space, only the retrieval results of four queries are shown from Figures A1 to A4. The test queries are randomly selected from among the 101 queries of the OHSUMED collection. In this correlation analysis, we show the results of query No.5, No.25, No.75, and No.95. However, other queries also produce similar results. These figures show that the correlation between document similarity and document granularity is low in all cases. Indeed, the Pearson's correlation coefficient value between the similarity scores and the granularity scores for all the documents is consistently low ($<\pm 0.40$) for each query. This correlation analysis reveals that document similarity and document granularity are two orthogonal dimensions of IR.

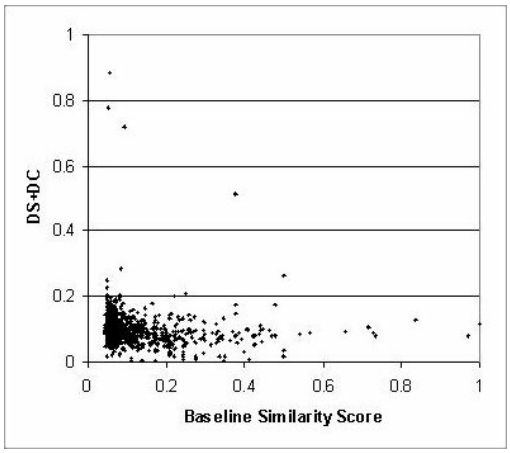

Fig. A1 – Correlation Analysis of Query 5

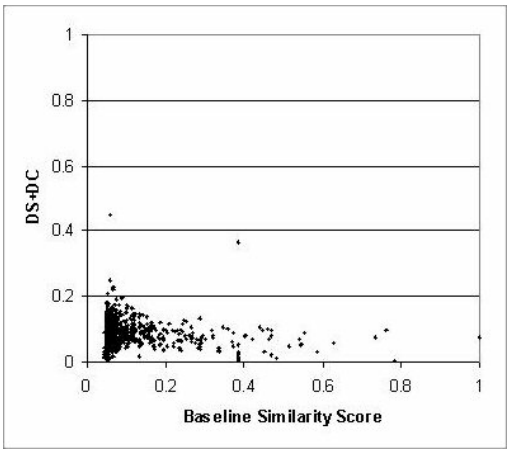

Fig. A2 – Correlation Analysis of Query 25

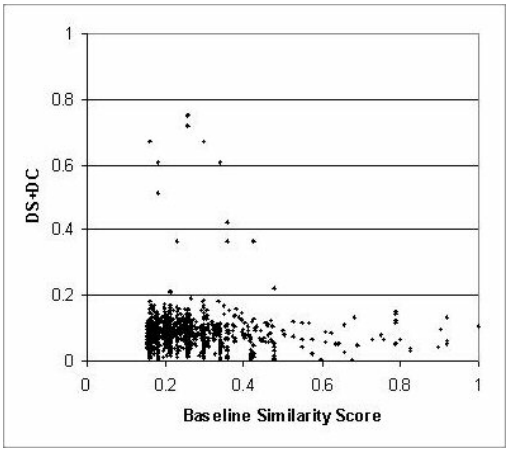

Fig. A3 – Correlation Analysis of Query 75

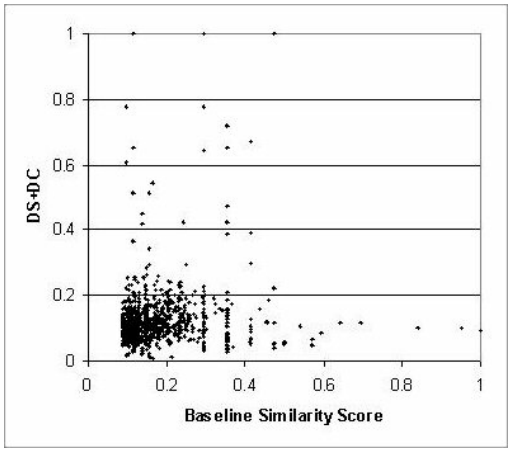

Fig. A4 – Correlation Analysis of Query 95

ACKNOWLEDGMENTS

The authors would like to thank four anonymous reviewers and the associate editor, Dr. Edward Fox, for their constructive comments that helped in improving the quality of the paper.

REFERENCES

ALLEN, R.B. AND WU, Y. 2002. Generality of texts. In *Proceedings of the 5th International Conference on Asian Digital Libraries*, Singapore, December 11-14, 2002, pp. 111-116.

ARONSON, A. R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the 2001 American Medical Informatics Association Annual Symposium*, pp.17–21.

BAILEY, P., CRASWELL, N., DE VRIES, A. P. AND SOBOROFF, I. 2007. Overview of the TREC 2007 enterprise track. In *Proceedings of The Sixteenth Text Retrieval Conference (TREC 2007)*, Gaithersburg, Maryland, USA, November 5-9, 2007. Available at: http://trec.nist.gov/pubs/trec16/papers/ENT.OVERVIEW16.pdf

BARGIELA, A. AND PEDRYCZ, W. 2008. Toward a Theory of Granular Computing for Human-Centered Information Processing. *IEEE Transactions on Fuzzy Systems*, 16(2): 320-330.

BEAULIEU, M., FOWKES, H., ALEMAYEHU, N. AND SANDERON, M. 1999. Interactive OKAPI at Sheffield TREC-8. In *The Eighth Text Retrieval Conference (TREC-8)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, November 17-19, 1999, pp. 17-19.

BELKIN, N.J., PEREZ CARBALLO, J., COOL, C., KELLY, D., LIN, S., PARK, S.Y., RIEH, S.Y., SVAGE-KNEPSHIELD, P. AND SOKORA, C. 1998. Rutgers' TREC-7 interactive track experience. In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 275-283.

BELKIN, N.J., HEAD, J., JEGN, J., KELLY, D., LIN, S., PARK, S.Y., COOL, C., SAVAGE-KNEPSHIELD, P. AND SIKORA, C. 1999. Relevance feedback versus local context analysis as term suggestion devices: Rutgers' TREC-8 interactive track experience. In *The Eighth Text Retrieval Conference (TREC-8)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 565-574.

BHATIA N., SHAH N.H., RUBIN D.L., CHIANG A.P., AND MUSEN M.A. 2009. Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. In *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*. Available at: https://bmir.stanford.edu/file_asset/index.php/1349/BMIR-2008-1332.pdf.

BODNER, R.C. AND CHIGNELL, M.H. 1998. CLICKIR: text retrieval using a

dynamic hypertext interface. In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 573-582.

BUYUKKOKTEN, O., KALJUVEE, O., GARCIA-MOLINA, H., PAEPCKE, A., AND WINOGRAD, T. 2002. Efficient Web browsing on handheld devices using page and form Summarization. *ACM Transactions on Information Systems*, 20(1):82-115.

CARBONELL, J. AND GOLDSTAIN, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, August 24-28, pp. 335-336.

DUMAIS, S., PLATT, J., HECKERMAN, D., AND SAHAMI, M. 1998. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, Bethesda, Maryland, November 3-7, pp. 148-155.

FONSECA. F., EGENHOFER, M., DAVIS, C., AND CAMARA, G. 2002. Semantic Granularity in Ontology-Driven Geographic Information Systems. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):121-151.

FLEISS, J.L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

FULLER, M., KASZKIEL, M., KIM, D., NG, C., ROBERTSON, J., WILKINSON, R., WU, M. AND ZOBEL, J. 1998. TREC 7 Ad Hoc, speech, and interactive tracks. In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 465-474.

FULLER, M., KASZKIEL, M., KIMBERLEY, S., ZOBEL, C., NG, J., WILKINSON, R. AND M, W. 1999. The RMIT/CSIRO Ad Hoc, Q&A, Web, interactive and speech experiments at TREC-8. In *The Eighth Text Retrieval Conference (TREC-8)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 549-564.

GAN, G., MA, C. AND WU, J. 2007. *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Alexandria, VA.

GEY, F., JIANG, H., CHEN, A. AND LARSON, R.R. 1998. Manual queries and machine translation in cross-language retrieval and interactive retrieval with Cheshire II at TREC-7. In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 527-540.

GRANKA, L., JOACHIMS, T. AND GAY, G. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, pp. 478-479.

HAVELIWALA, T.H. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784-796.

HE, B. AND OUNIS, I. 2004. Inferring query performance using pre-retrieval predictors. In *11th Symposium on String Processing and Information Retrieval*, Padova, Italy, October 5-8, LNCS, Springer-Verlag, pp. 43-54.

HERSH, W., BUCKLEY, C., LEONE, T.J. AND HICKAM, D. 1994. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval*, New York, NY, USA, Springer-Verlag pp. 192-201.

HERSH, W., PRICE, S., KRAEMER, D., CHAN, B., SACHEREK, L. AND OLSON, D. 1998. A large-scale comparison of boolean vs. natural-language searching for the TREC-7 interactive track. In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 491-500.

HERSH, W. 1999. TREC-8 Interactive Track Report. In *The Eighth Text Retrieval Conference*, Gaithersburg, Maryland, November 17-19, pp. 57-64.

HERSH, W., TURPIN, A., PRICE, S., KRAEMER, D., CHAN, B., SACHEREK, L. AND OLSON, D. 1999. Do batch and user evaluations give the same results? An analysis from the TREC-8 interactive track. In *The Eighth Text Retrieval Conference (TREC-8)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 17-24.

HERSH, W., COHEN, A.M., ROBERTS, P., REKAPALLI, H.K. 2006. TREC 2006 Genomics Track Overview. In *The Eighth Text Retrieval Conference (TREC-15)*, E.M. VOORHEES and L. P. Buckland Eds., Gaithersburg, Maryland, USA, pp. 52-78.

HO, J. AND TANG, R. 2001. Towards an optimal resolution to information overload: an infomediary approach. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, Boulder, Colorado, USA, September 30 - October 03, pp. 91-96.

LAGERGREN, E. AND OVER, P. 1998. Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ACM Press, New York, NY, USA, pp. 164-172.

LARSON, R.R. 1999. Berkeley's TREC-8 interactive track entry: Cheshire II and Zprise. In *The Eighth Text Retrieval Conference (TREC-8)*, E.M. VOORHEES

and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 613-622.

LAU, R.Y.K., SONG, D., LI, Y., CHEUNG, C.H., AND HAO, J.X. 2009a. Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):800-813.

LAU, R.Y.K., LAI, C.L., AND LI, Y. 2009b. Mining Fuzzy Ontology for a Web-Based Granular Information Retrieval System, *Proceedings of the Fourth International Conference on Rough Set and Knowledge Technology*, Gold Coast, Australia, 14-16 July 2009, Volume 5589 of Lecture Notes in Computer Science, Springer-Verlag, pp. 239-246.

LAU, R.Y.K., BRUZA, P.D., AND SONG, D. 2008. Towards a Belief Revision Based Adaptive and Context Sensitive Information Retrieval System. *ACM Transactions on Information Systems* 26(2):8.31-8.38.

LEACOCK, C. AND CHODOROW, M. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet*, *an Electronic Lexical Database*, Fellbaum, C. Ed., Cambridge, MA, MIT Press, 1998, pp. 265–283.

LI, Y., BANDAR, Z.A., AND MCLEAN, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871-882.

LIU, Y., ZHANG, B., CHEN, Z., LYU, M.R. AND MA, W.Y. 2004. Affinity rank: a new scheme for efficient Web search. In *Proceedings of the Thirteenth World Wide Web Conference* ACM, New York, USA, pp. 338-339.

LIU, X. AND CROFT, W.B. 2002. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, November 4-9, pp. 375-382.

LIU, Y., YANG, Y., AND CARBONELL, J. 2002. Boosting to correct inductive bias in text classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, November 4-9, pp. 348-355.

MORRIS, J. AND HIRST, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* Cambridge, MA, USA: MIT Press, 17:21-48.

MOWSHOWITZ, A. AND KAWAGUCHI, A. 2002. Bias on the Web. *Communications of the ACM*, 45(9): 56–60.

OGDEN, W., DAVIS, M. AND RICE, S. 1998. Document thumbnail visualizations for rapid relevance judgements: when do they pay off? In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 528-534.

OVER, P. 1997. TREC-6 interactive track report. In *The Sixth Text Retrieval*

Conference, Gaithersburg, Maryland, November 19-21, 1997, pp. 73-82.

OVER, P. 1998. TREC-7 interactive track report. In *The Seventh Text Retrieval Conference*, Gaithersburg, Maryland, November 9-11, 1998, pp. 65-72.

PAGE, L., BRIN, S., MOTWANI, R. AND WINOGRAD, T. 1998. The PageRank citation ranking: bringing order to the web. Technical Report, Stanford InfoLab. Available at: http://ilpubs.stanford.edu:8090/422/.

PENG, F., SCHUURMANS, D., AND WANG, S. 2004. Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval*, 7(3-4): 317-345.

PLACHOURAS, V., CACHEDA, F., OUNIS, I. AND VAN RIJSBERGEN, C.J. 2003. University of Glasgow at the Web track: dynamic application of hyperlink analysis using the query scope. In *Proceedings of the 12th Text Retrieval Conference TREC 2003*, Gaithersburg, Maryland, November 18-21, pp. 636-642.

PORTER, T. 1980. An algorithm for suffix striping. *Program*, 14(3): 130-137.

PONTE, J. AND CROFT, B. 1998. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275-281.

PRICE, S.L., NIELSEN, M.L., DELCAMBRE, L.M.L., AND VEDSTED, P. 2007. Semantic components enhance retrieval of domain-specific documents. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, Lisbon, Portugal, pp. 429-438.

RANSDELL, J. 1966. Charles Peirce: the idea of representation. Ph.D. dissertation, Columbia University, New York, USA, Retrieved March 23, 2010, from Dissertations & Theses: A&I. (Publication No. AAT 6709367).

RESNIK, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 448-453.

ROBERTSON, S.E. 1997. The probability ranking principle in IR. In *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., pp. 281-286.

ROBERTSON, S.E., WALKER, S. AND BEAULIEU, M. 1998. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In *The Seventh Text Retrieval Conference (TREC-7)*, E.M. VOORHEES and D.K. HARMAN Eds., Gaithersburg, Maryland, USA, pp. 253-264.

ROUSSINOV, D. G., AND CHEN, H. 2001. Information Navigation on the Web by Clustering and Summarizing Query Results. *Information Processing and Management*, 37(6):789 – 816.

SALTON, G., WONG, A. AND YANG, C.S. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18: 229-237.

SALTON, G. AND BUCKLEY, C. 1988. Term-weighted approaches to automatic text retrieval. *Information Processing and Management*, 24(5): 513-523.

SALTON, G. 1990. Full Text Information Processing Using the Smart System. *IEEE CS Technical Communications on Database Engineering Bulletin*, 13(1): 2-9.

SALTON, G. 1991. Developments in automatic text retrieval. *Science*, 253(5023): 974-980.

SANTAELLA, L. 2003. What is a symbol. *Semiotics, Evolution, Energy, and Development* 3: 54-60.

SHEPARD, R. 1987. Towards a universal law of generation for psychological science. *Science*, 237:1317-1323.

SWAN, R.C. AND ALLAN, J. 1998. Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* ACM Press, New York, NY, USA, pp. 173-181.

VAN RIJSBERGEN, C.J. 1979. *Information retrieval*. London; Boston: Butterworths.

WANG, M. AND SI, L. 2008. Discriminative probabilistic models for passage based retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 419-426.

YAO, J.T. 2005. Information granulation and granular relationships. In *Proceedings of the 2005 IEEE International Conference on Granular Computing*, pp. 326-329.

YAO, Y.Y. 2002. Information retrieval support systems. In *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, pp. 773-778.

YAN, X., SONG, D., AND LI, S. 2006. Concept-based document readability in domain-specific information retrieval. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 540-549.

YANG, K., MAGLAUGHLIN, K.L., MEHO, L., SUMNER, R.G, JR. 1998. IRIS at TREC-7. In *The Seventh Text Retrieval Conference (TREC-7)*, Gaithersburg, Maryland, USA, pp. 555-566.

YANG, K., MAGLAUGHLIN, K.L. AND IRIS, J. 1999. TREC-8. In *The Eighth Text Retrieval Conference (TREC-8)*, Gaithersburg, Maryland, USA, pp. 645-656.

ZADEH, L.A. 1979. Fuzzy sets and information granularity. In *Advances in Fuzzy Set Theory and Applications*, M. Gupta, R.K. Ragade, R.R. Yager (eds),

North-Holland Publishing Company, pp. 3-18.

ZAKOS, J., VERMA, B., LI, X. AND KULKARNI, S. 2003. Intelligent encoding of concepts in Web document retrieval. In *Proceedings of the fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03)*, pp. 72.

ZHAI, C. 2002. Risk minimization and language modeling in text retrieval dissertation abstract. *SIGIR Forum 36*, pp. 100-101.

ZHAI, C., COHEN, W.W. AND LAFFERTY, J. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 10-17.